How Informational Teleosemantics Works

Towards a realist theory of content

by

Johannes Heemskerk

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Philosophy

Philosophy Department

University of Warwick

September, 2024

Contents

List of Figures

List of Tables

Acknowledgements

I want to thank a great number of people who have helped me throughout my thesis. These include, but are not limited to:

Those who attended the various conferences at which I presented my work. I received particularly helpful feedback from Madelaine Angelova-Elchinova, Hilary Kornblith, Edouard Machery, Gualtiero Piccini, Manolo Martinez, and a host of names I have forgotten.

Those friends and colleagues within my department who I bored with information theory: Nadine Elzein, Hemdat Lerman, Christoph Hoerl, Tom Crowther, Lucy Campbell, Richard Moore, and especially my good friends Sean Manners, Dino Jakusic, Oscar North, Clarissa Muller, Brigid Evans and Bruna Picas I Prats who provided me with much joy during the PhD and made the whole thing bearable.

Special mention to Carl Frietsch, who has been my friend for many years, and to whom I owe a great deal of intellectual and emotional support.

Many thanks to Sabina Pachlopníková, my wonderful partner, who made the especially difficult last few months not only bearable but positively uplifting.

I would also like to thank Rachel Jose, who supported me in a number of ways throughout my studies.

Also to my parents and sister, for so much help in a number of ways.

Finally (though I have probably forgotten so many people), I want to thank Stephen Butterfill. Simply the very best supervisor I could have ever hoped for, and the person who inspired me to take up the kind of philosophy I have conducted in this thesis.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Abstract

Representations appear to play a central role in cognitive science. Capacities such as face recognition are thought to be enabled by internal states or structures representing external items. However, despite the ubiquity of representational terminology in cognitive science, there is no explicit scientific theory outlining what makes an internal state a representation of an external item. Nonetheless, many philosophers hope to uncover an implicit theory in the scientific literature. This is the project of the current thesis. However, all such projects face an obstacle in the form of Frances Egan's argument that content plays no role in scientific theorising. I respond that, in some limited regions of cognitive science, content is crucial for explanation. The unifying idea is that closer attention to the application of information theory in those regions of cognitive neuroscience enables us to uncover an implicit theory of content. I examine the conditions which must be met for the cognitive system to be modelled using information theory, presenting some constraints on how we apply the mathematical framework. For example, information theory requires identifying probability distributions over measurable outcomes, which leads us to focus specifically on neural representation. I then argue that functions are required to make tractable measures of information, since they serve to narrow the range of possible contents to those potentially explanatory of a cognitive capacity. However, unlike many other teleosemanticists, I argue that we need to use a non-etiological form of function. I consider whether non-etiological functions allow for misrepresentation, and conclude that they do. Finally, I introduce what I argue is the implicit theory of content in cognitive neuroscience: maxMI. The content of a representation is that item in the environment with which the representation shares maximal mutual information.

Dedication

To loved ones, present or absent.

Chapter 1

Introduction

> *It is Ambition enough to be employed as an Under-Labourer in clearing Ground a little, and removing some of the Rubbish, that lies in the way to Knowledge.*

John Locke, Epistle to the Reader, *An Essay Concerning Human Understanding*

1.1 Background

Minds provide their users with a range of cognitive capacities - abilities which require the exercise of mental processes. Cognitive science is an attempt to understand and explain those capacities using broadly scientific methodology. The explanations offered by cognitive science defer to things called representations. Each representation is thought to have a special relation to a particular thing external to itself, that thing being known as the 'content' of that representation.

Explanations in cognitive science also involve the positing of computations performed over those representations (e.g. [Friedenberg et al., 2021, p3]). These computations are

thought to respect the content of the representations, in the sense that internal processing is partly determined by the thing which the representation represents. This is how the cognitive system is able to enact cognitive capacities which are directed upon those external items.

For example, humans and other animals possess the capacity to recognise familiar faces. We can do this in a range of conditions - the face can be side-on, dimly lit, partially occluded, and so on. A leading theory in cognitive science suggests that facial recognition is achieved, in part, by the system performing computations over perceptual representations elicited by a face. This processing results in the mapping of those perceptual representations to an internally-stored 'face-space' representation (see [O'Toole, 2011] for a review). Mapping to the face-space allows the perceived face to be represented on a standard template, allowing comparison with stored representations of familiar faces, despite variations in the input. This explanation relies on representational content in a number of ways. For example, the content of the perceptual representation determines its processing; whether the perceptual representation is mapped to a face-space or not depends on whether its content is appropriately related to the content of the face-space.

However, there is no consensus in cognitive science about how content is to be attributed to a representation. In fact, despite explanations relying on representational content, we currently have no general theory of content which tells us, for any given representation, what makes it the case that some item(s) is/are the content of that representation. This also means there is no explicit procedure for determining content which could be used across the various disciplines which make up cognitive science.

In this thesis I will argue that while there is no such explicit theory of content in

cognitive science, there is an *implicit* theory which is operative in certain domains of cognitive science. Across much of the field, the methodologies employed to discover content - such as spike-triggered averages, conditional mutual information, maximally informative dimensions - testify to an implicit theory of the relationship between contents and representations. For those within cognitive science who find that the implicit theory serves their explanatory needs, it can, once explicit, facilitate the systematisation of content attributions for those projects.

1.2 Research questions

The following three questions guide the thesis:

1. Is there an implicit theory of content in cognitive science?

2. How can we discover the implicit theory of content?

3. What is the implicit theory of content?

Each question relies, at least in part, on an answer to the other two. It will be difficult to establish that there is an implicit theory without showing what it is. It will be impossible to show what the theory is without knowing how to discover it. It will be difficult to know how to discover an implicit theory unless we know whether there is one.

In order to make some headway, in chapter three I consider an argument, due to Frances Egan, which would have us conclude that there is no implicit theory of content in cognitive science (e.g. [Egan, 2014]). Egan presents an argument which, in effect, spells out necessary criteria for content having an explanatory role within cognitive science.

Egan argues that if the use of content does not meet these necessary criteria, then content attributions are not governed by an implicit theory.

I will endorse Egan's criteria but, *conta* Egan, I will show that content use can meet them, provided we are selective about the studies we use to investigate content attribution. I spell out three further criteria we should use to ensure that content plays an explanatory role within any study we investigate (section 3.5).

Following chapter three, we will have gone some way to answering the first and second research questions. We will know, at least, that an implicit theory of content is possible. We will also have narrowed our search considerably; using our set of constraints we will know which regions of cognitive science *could* harbour an implicit theory of content, and which we can ignore. I go on to employ the methodology spelled out below to continue the search for an implicit theory of content within the promising regions of cognitive science.

1.3 Method

My approach will be largely similar to, among others, that of both Ramsey (e.g. [Ramsey, 2007]) and Orlandi (e.g. [Orlandi, 2020]), who analyse scientific practice in order to understand how representational explanation is used within that practice. As Orlandi writes:

> I propose that we look at what mental representations are by looking at how they have been used in these disciplines [i.e. the cognitive sciences]. In this respect, I take philosophers interested in the notion of mental representation to be akin to those philosophers and historians of science more generally

who investigate the nature of scientific posits by looking at scientific practice. [Orlandi, 2020, p101]

As I will be understanding and employing it, the methodology outlined by Orlandi is an application of scientific reasoning to science itself. In this sense, "looking at scientific practice" involves, first, providing a hypothesis for the theory of content which guides scientific practice. Second, generating predictions about how scientific practice would proceed were the hypothesised theory operative, in contrast with predictions about how the science would proceed were another theory operative. Finally, studying scientific practice in order to ascertain which predictions are borne out.

When attempting to extract information - in this case, about the implicit theory of content in cognitive science - it is crucial to have some structuring hypothesis with which to interpret the evidence - in this case, the cognitive science literature. Of course, we must be careful to avoid confirmation bias, so throughout I will raise complications for the theory which test whether the guiding hypothesis really accounts for scientific practice.

In order to generate the hypothesis, I employ insights from philosophy. In chapter two, I provide a literature review of existing philosophical theories of content, and situate my own project among them. I primarily develop a line of thinking, beginning with Fred Dretske, which suggests that the content of a representation is determined by both its informational link with some environmental item (e.g. [Dretske, 1981]) and its functional role within a wider system (e.g. [Dretske, 1994]). I use this guiding hypothesis to specify research question three:

1. What is the implicit theory of content?

i. Which type of information link is relevant for content determination?

ii. Which type of function is relevant for content determination?

To answer (i), I first show, in chapter four section 4.4, that Claude Shannon's mathematical model of information theory [Shannon, 1948], specifically, provides what I call the 'background theoretical framework' of content attribution in the relevant regions of cognitive science. I then consider, in chapter seven section 7.2.1, two models of the relevant information link - correlational information and mutual information - and argue, by way of comparing how each model best fits the methodology of cognitive science, that mutual information is the relevant information link.

To answer (ii), in chapter five I spell out precisely how information theory in Shannon's sense can be properly applied to the cognitive system. If we are to do so carefully, we must observe a number of further restrictions (e.g. section 5.3). I will argue, in chapter six, that the conditions for the application of Shannon's information theory require us to understand functions in the sense of Robert Cummins (e.g. section 6.5).

Each strand of the theory builds to make explicit an implicit theory of content as used in certain regions of cognitive science, as evidenced by both the methodology used and the commitments scientists must take on if they are to properly apply information theory (which they do). The theory, which I call maxMI, will be summarised in section 1.4.

There are some general constraints I will observe. Since content is supposed to be explanatory, I will focus on cases in which its explanatory role is most perspicacious. I will be following Nicholas Shea's dictum "externalist explanans, externalist explanadum" [Shea, 2018, p31]. I will primarily investigate studies for which a cognitive capacity clearly involves some item outside of the system itself in the external environment. I have

chosen to focus on two related categories of study: those pertaining to facial recognition and those pertaining to object recognition more generally. The cognitive capacity, recognition, is operationalised in various concrete ways, such as the "ability to assign labels (e.g. nouns) to particular objects, ranging from precise labels ('identification') to course [sic; coarse] labels ('categorization')." [DiCarlo et al., 2012, p416]. In this sense, recognition is a capacity humans, at least, have which is directed towards external items (faces or objects). Here, the explanatory value of content is most likely to be evinced, since the capacity involves the environment directly. So, if we hope to find an implicit theory governing content attribution, these studies are the most likely candidates in which to find one.

I will say more about how I understand content to be explanatory in chapter two section 2.5.

1.4 Conclusions

I conclude that, within certain regions of cognitive science, there is an implicit theory of content. The relevant regions are those areas of cognitive neuroscience, systems neuroscience, psychophysics, and any other region explicitly investigating neural representations which take into consideration the downstream decoding capacities of the cortex. More narrowly still, these areas must specify content using technical vocabulary which is sufficiently precise to enable the modelling of the external item as a random variable (section 5.2).

Decoding must be accounted for, since this ensures that a change in content makes

a change for the system itself. If downstream areas are insensitive to upstream changes, content cannot play the kind of explanatory role that other explanatory posits, such as the description of neural behaviour in terms of mathematical function computed, play. I spell this out in chapter three section 3.4.1 and chapter seven section 7.5.3.

Studies dealing with neural representation must be considered over those which deal with more folk psychologically-described states such as belief or intention, since the application of information theory requires that specifiable ranges with probabilities over their values be identified. I spell out this argument in chapter five section 5.5.

Technical vocabulary must be used. First, this facilitates the modelling of the external item as a random variable. Second, it ensures that content can be operationalised, meaning that content attributions can be tested via standard empirical methods. I spell this out in chapter three 3.4.2.

I conclude that the the implicit theory is what I call **maxMI**:

> **maxMI**: E_x is the content of R iff R shares mutual information with each of a set of items, E_{1-n}, of which E_x is a member, and R and E_x have maximal mutual information relative to the rest of E_{1-n}.

Where E_x is some item external to the representation, and R is a representation. In broad outline, the above provides the theory of content I argue is implicit in cognitive neuroscience. However, additional constraints must be added for reasons which will be presented throughout the thesis:

1. R must be modelled as an iRV with outcomes constrained by values usable for downstream systems.

- See sections 5.3, 6.2, 7.5.3.

2. E_x must be modelled as an eRV with outcomes constrained by values discriminable by sensory interfaces.

    - See sections 2.4.1, 5.4, 5.4.1.

3. The set E_{1-n} must be delimited by the C-function of the subsystem containing R.

    - See sections 5.4.2, 6.2, 6.5.

Where "iRV" stands for "internal random variable", "eRV" for "external random variable", and "C-function" for the type of functions introduced by Robert Cummins. I provide these conditions here for ease of reference - I hope the reason for them will become clear as one reads through the thesis. In short, each condition, taken together with maxMI, delivers content attributions which ensure that the content of a representation is content for the *system itself*; content a change in which results in a change for the system.

1.4.1 Toy example

In this section I introduce a simplified example to illustrate how maxMI picks out content in practice. The toy example is not supposed to show how cognitive neuroscientists actually determine content. In chapter seven, section 7.4, I outline their actual methods, and how they presuppose maxMI. Instead, the example is intended to make the theory more intuitive by showing how it can be applied to attribute contents under very simplified conditions.

Imagine we are developing a theory about how people recognise faces. Imagine that we have isolated a single neuron, R, which we believe to be responsible for facial recognition. When active, which we can label R_{on}, this neuron provides us with a representation which allows us to discriminate between faces. What we want to know is precisely what R_{on} represents - what is its content? Imagine we have three hypotheses; R_{on} represents either:

X: The separation between eyes (EyeSep)

Y: The separation between ears (EarSep)

Z: The width of the mouth (MouthWid)

In intuitive terms, we hypothesise that people recognise faces *either* according to how far apart the eyes are, or how far apart the ears are, or how wide the mouth is (at resting) on the target face. Of course, this won't work in reality: multiple different people likely have eyes (or ears) the same distance apart (or mouths of the same width). EyeSep alone is, in actuality, a very poor way to discriminate between faces. We will simplify and assume this is possible.

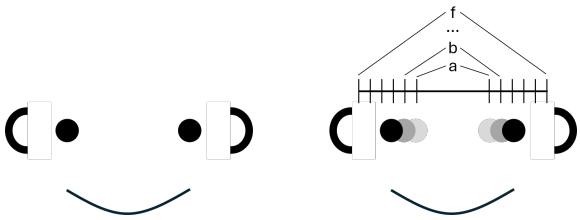


Figure 1.1: A face with possible distances between the eyes labelled a to f.

As in figure 1.1, we can assume for simplicity that there are a discrete number of possible separation and width values. The eyes can be a distance apart, up to f distance

apart. Similarly for the ears, which can be from g distance apart to l distance apart. The mouth can be m units wide up to r units wide.

X		Y		Z	
a	0.22	g	0.14	m	0.03
b	0.14	h	0.09	n	0.05
c	0.19	i	0.26	o	0.27
d	0.15	j	0.13	p	0.31
e	0.06	k	0.08	q	0.20
f	0.24	l	0.30	r	0.14

Table 1.1: Values a-f for X (EyeSep), g-l of Y (EarSep), m-r of Z (MouthWid) with corresponding probabilities.

Table 1.1 shows the corresponding probability ranges for each possible position of the eyes, ears, and mouth. Imagine we obtain these by performing research to discover the probability distribution across all faces (given typical constraints on conducting such research). In reality, this is dummy data.

Imagine that R can be in two states, R_{on} or R_{off}. The probability of $\mathrm{R}_{on} = 0.1$ and the probability of $\mathrm{R}_{off} = 0.9$. We now have four random variables. X, Y, and Z are our eRVs - the random variables which model items in the environment. In this case, they model specific distances between each feature or the width of the mouth. R is our iRV with R_{on} and R_{off} as outcomes.

We can now invent some conditional probabilities. For example, we can invent $p(R_{on}|a)$, $p(R_{on}|g)$, and $p(R_{on}|m)$. In fact, we need to invent these values for each condition a-r for both R_{on} and R_{off}. Imagine we conduct some empirical tests to discover these conditional probabilities - we find the probability that R is active when presented with various separation and width values. In reality, this is more dummy data. I present a snapshot to give an indication of the type of values which are relevant in table 1.2.

$p(R_{on}\|a)$	$p(R_{on}\|g)$	$p(R_{on}\|m)$
0.545	0.357	0.667

Table 1.2: Dummy values for relevant conditional probabilities

As will become relevant later (section 7.2.1), note that R_{on} is correlated with each condition. As a result, R has mutual information with each of X, Y and Z. What we need to find now is the eRV with which the iRV has maximal mutual information. To measure amounts of mutual information we use this formula:

$$I(X,Y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} p(x_i, y_i) log \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \tag{1.1}$$

Using our conditional probability and marginal probability dummy data, we can use Bayes' theorem to derive joint probabilities, and insert them into equation 8.5 to obtain mutual information values as given in table 1.3.

$I(X,R)$	$I(Y,R)$	$I(Z,R)$
1.147	0.230	0.451

Table 1.3: Calculated MI values

According to maxMI, provided we have correctly chosen our eRVs and iRVs, X models the content of R - which is therefore $\langle EyeSep \rangle$[1]. In terms of answering our initial explanatory question, this means that eye separation is how we are able to recognise faces.

[1]Throughout the thesis I follow the standard practice of placing content ascriptions within chevrons.

1.4.2 Summary of maxMI

As pointed out by Randy Gallistel, mutual information provides a measure of the precise strength of correlation between two random variables, making it the most promising model of the content-determining information link. I spell this out in chapter seven section 7.2.1. Importantly, *maximal* mutual information provides content with its explanatory power in virtue of providing a model of the information "available" to the system itself. I spell this out in chapter seven section 7.5.

Altogether, maxMI unifies the various methodological and theoretical commitments taken on by researchers within the relevant areas of cognitive science. In chapter two, I spell out the relations between my own approach and that of other philosophers. The theory I propose differs significantly in several respects from what has come before, most notably by rejecting etiology and by embracing Shannon's information theory in earnest.

In the conclusion I consider what is gained by making the implicit theory of content explicit. I propose that it may help with projects which aim to discover representational content in such a way that can be integrated with work in other regions of cognitive science. For example, in comparative psychology, comparing human and non-human cognitive capacities benefits from tracing differences in the kind of content isolated by maxMI.

I also conclude that much of cognitive science may be using content in a way which is not driven by an implicit theory. To use Egan's term, in many cases content may be a "gloss" on the explanation itself - a way of making a theory comprehensible to a reader, but not essential to the explanation. The implicit theory may or may not be helpful in such areas of cognitive science, but it should encourage greater engagement between

fields if adopted where appropriate.

1.4.3 Translating the mathematical model

Information theory provides a mathematical model of the content-determining relation. However, in chapter four I consider how we can map the mathematical model to empirical reality. Described in a non-mathematical way, maxMI is the theory that content is determined by way of the functional connectivity of a representation to initial sensory receptors on the one side, and downstream capacities on the other.

According to the theory, ranges are established in the brain by way of functional connections to upstream areas (receptive fields) which are connected to receptors (their receptive fields) which pick up incoming information (see section 5.4). A range in the environment, for example a range of electromagnetic frequencies, has a corresponding voltage range along axons of ganglion cells - corresponding in the sense that it is a C-function of the ganglion cell to transmit information about light frequency (see the glossary for a definition of C-function). These ranges are transformed, by way of invariance mechanisms (see section 5.4.1), into more complex internal ranges which correspond to the more complex external ranges which produced the electromagnetic frequencies.

Information theory models the relationship between a value of the internal range and a value of the external range. Maximal mutual information picks out that item in the world which a value of the internal range is most informative about. From our perspective, focusing on the downstream areas which use the information, this means that downstream areas will act as if that value of the external item is present.

1.5 Definitions

A full glossary is available at the end of the thesis. Here, for the sake of intelligibility, I highlight a few key terms and phrases which will come up frequently in the thesis.

As I will use the phrase, a '**theory of content**' is a theory which allows one to specify, once the relevant empirical facts are known, what the content of any given representation is.

As I use the term, a '**representation**' is a mediating state, structure or process between an input and an output (e.g. upstream receptors and downstream systems) with content. In this definition I follow Shea, who argues that what a representation *is* must be connected to a theory of content [Shea, 2018, p10].

As I use the term, '**content**' is an item external to the representation, either in the 'external world' or internal to the system (i.e. a representation can represent some other aspect of the system). I follow Neander's use of the term 'content' to specify an *item* itself [Neander, 2017b, p15]. This is in contrast with other uses of the term according to which 'content' only specifies something mental - such as a mental image. In my sense, calling E the content of representation R is like calling S the aunt of A: they are related in a specific way.

I will use the term '**item**' to refer to things in the world aiming to be neutral on their specific properties. I follow Shea in intending this usage to be "neutral about what should count as an item. It could be a particular object, [...] it could be a collection of objects or a type of object [...]. It could also be a process or a type of process" [Shea, 2018, p76n]. Many types of item or items, provided they meet the criteria set out in (1) of maxMI, can be the content of a representation.

1.6 Scope of the thesis

The maxMI theory is not stand-alone. If the theory is to make any claim to general veracity, it must be the case that I have accurately captured the relevant region of scientific practice, and that scientific practice itself produces adequate explanations of cognitive capacities. As Neander remarks, whether the theory isolates anything "real" "is conditional on these sciences not being on the wrong track in relevant respects" [Neander, 2017b, p78]. It is not my aim to show that they are. If they are on the wrong track, maxMI will be more or less useless. However, I am an optimist. To my knowledge, there is no reason to suspect that the basic approach of these studies is fundamentally wrong. Quite the opposite, they largely represent the most state-of-the-art and sophisticated approaches which fix some issues with previous work (see section 4.4).

Given the focus on neural representation, maxMI is of necessity "modest" in Peter Godfrey-Smith's sense [Godfrey-Smith, 1998]: it aims to capture the content of relatively low-level representations, those Neander characterises as "nonconceptual or preconceptual sensory-perceptual representations (perhaps together with a relatively small set of core concepts)" [Neander, 2017b, p10]. However, I make some comments in the conclusion (section 8.3) and in chapter five 5.4.1 on the scaling-up problem faced by theories of content, and whether maxMI has any resources to address it. As I will spell out, I think there is some reason to be optimistic here, too.

Generally, I will not be explicitly aiming to address traditional philosophical desiderata on representational content, except in cases in which such a discussion is directly relevant to the analysis (e.g see section 6.6.2). As mentioned above, the theory is primarily aimed at addressing the questions of section 1.2. Whether there is an implicit theory of

content in cognitive science, and what it is, can be detached from the question of whether such a theory meets philosophical desiderata. Of course, if cognitive science is to be on sure theoretical footing, it should meet such desiderata (except in cases in which we have independent reason to question the desiderata themselves).

However, I do address what I take to be some central philosophical concerns. For example, I consider the question of indeterminacy. I provide a way of understanding the relevancy of determinate content for scientific theorising in chapter three (section 3.4) and chapter seven (section 7.3). I will conclude that indeterminacy ought not to concern us except insofar as it renders content indeterminate with respect to providing an adequate explanation of a cognitive capacity. maxMI meets the criteria I set out for this, so should not be seen as producing problematically indeterminate content ascriptions. If, for example, a range of items each have the same mutual information value with a representation, I consider these items to be in the "content profile" of the representation, evincing a case of "natural indeterminacy". However, in the conclusion I highlight some outstanding questions about indeterminacy resulting from maxMI (section 8.4.2).

Additionally, I provide an answer to the question whether contents should be spelled out in terms of "high-church" or "low-church" properties (see section 2.5). I answer that which answer one gives will depend on the explanatory role one attributes to content.

I also consider the problem of misrepresentation. In chapter six (section 6.6), I argue that the functions relative to which content is determined are able to malfunction or be dysfunctional, securing the possibility that the content attributed according to maxMI can misrepresent. I claim that this is a pressing concern due to the prevalent paradigm in cognitive science involving using dysfunctional patients to infer facts about the cognitive

systems of 'normally' functioning individuals.

I also provide a brief discussion of "swampman" cases - thought experiments in which an individual with the very same physical structure as an existing person is created instantaneously by a freak bolt of lightning in a swamp. I argue that swampman is a serious consideration - it forces us to consider the explanatory value of our theory - but it is not a problem for maxMI, which relies on non-etiological functions (section 6.7).

What I will explicitly not consider is whether the content attributions made according to maxMI accord with our intuitions. I agree with Shea that "intuitions have little probative value for our kind of project" [Shea, 2018, p21].

1.7 Conclusion

The thesis aims to establish a relatively modest claim: in some regions of cognitive science there is an implicit theory of content. Moreover, the implicit theory of content isolates contents which are genuinely explanatory of cognitive capacities. Nonetheless, the road to such a claim involves more twists and turns than one may initially think. I hope to make the path as clear as possible.

The project requires some initial guidance. So, I will first turn to existing naturalistic accounts of content. Specifically, I will turn to what I consider to be the most promising research project: informational teleosemantics. I start the thesis by considering the work which has come before, and how it informs the current project.

Chapter 2

Informational Teleosemantics

2.1 Introduction

Cognitive science frequently uses informational terminology when describing representation. The cognitive system is often understood as an information processing system. I will argue in chapter three (section 4.2) that this is not loose talk; Shannon's information theory is the background theoretical framework of cognitive science. As such, a branch of philosophy known as informational teleosemantics is a *prima facie* promising framework to theorise about content in cognitive science. As I suggest in sections 2.4, 2.5, and 2.6.2, several informational teleosemanticists have made significant advances in developing a theory of content consistent with its use in cognitive science. Throughout the thesis, I aim to show how the framework can capture some fundamental commitments of the implicit theory of content in cognitive science. In this chapter, I outline informational teleosemantics with the aim of setting out some key theoretical concepts which will be

used in the remainder of the thesis.

Informational teleosemantics is, broadly, the view that representational content is determined by (i) an informational link between a representation and an item in the environment and (ii) the function of the relevant representational system[1]. Theories within the framework vary with respect to the way in which (i) and (ii) are spelled out. For example, how we should think of the relevant environmental item[2], the relevant type of informational link[3], the relevant type of functions[4], the explanatory aims of a theory of content[5], and how to answer a host of problems faced by any theory of content, such as indeterminacy, realism versus instrumentalism, pragmatism versus naturalism, the problem of misrepresentation, and more - see [Schulte and Neander, 2022] for an overview.

In this chapter I will first introduce informational teleosemantics by outlining the views of two major contributors - Fred Dretske and Ruth Millikan. Roughly, Dretske introduced information and Millikan functional teleology. I will then consider a crucial recent development in teleosemantics known as the "explanatory turn" (a phrase coined by Peter Schulte [Schulte, 2023]). In recent years, theorists have become increasingly interested in the question whether content ascriptions have any scientific explanatory value. I introduce a central theorist, Karen Neander, who distinguishes two types of explanatory project - drawing on Mayr's proximate/ultimate distinction [Mayr, 1961] - and considers which project content ascriptions serve. Neander argues that content ascriptions can serve proximate explanations. I agree with Neander, and intend maxMI to

[1] **Informational**: the relation between representation and content is spelled out in terms of information as in (i); **teleo**: functions as in (ii); **semantics**: to do with a theory of content.

[2] For example, section 2.6.1.

[3] For example, sections 2.6.3, 7.2

[4] For example, sections 2.3.1, 6.4

[5] For example, section 2.5

address the proximal explanatory project. However, I argue that Neander's own theory of content fails to address the proximal explanatory project she aims for.

I end by outlining how contemporary informational teleosemantic theories, such as those of Peter Schulte, Manolo Martinez, and Marc Artiga, are typically aimed at providing content attributions which feature in ultimate explanations. I then introduce Gualtiero Piccinini, who explicitly aims to provide content attributions for proximal explanations and provides helpful guidance in doing so. However, I suggest that attending to scientific practice complicates what might initially seem like a straightforward way of attributing content in scientific explanations. The remainder of the thesis is an attempt to outline the complex assumptions behind scientific practice, aiming to uncover the implicit theory of content in cognitive science.

2.2 Fred Dretske and information

Fred Dretske brought information theory into the philosophical mainstream. He demonstrated the power of Shannon's mathematical framework for modelling the content-determining relation between a representation and an external item. He made a number of foundational contributions to informational teleosemantics[6]. Many of his insights stand the test of time, and are reflected in the implicit theory of contemporary cognitive science, as we will see throughout the thesis.

[6]Although functions play no explicit role in his early work, in 1994's *If You Can't Make One, You Don't Know How It Works* he argued that functions were required for misrepresentation.

2.2.1 Informational content

I will focus primarily on Dretske's notion of 'informational content'. Dretske intends 'informational content' to provide the foundation for a "semantic theory of information" [Dretske, 1981, p64], delivering the "underlying informational structure" of "intentionality" - the so-called 'aboutness' of mental representation [Dretske, 1981, p76].

Informational content is given by:

> A signal *r* carries the information *s* is *F* = The conditional probability of *s*'s being *F*, given *r* (and *k*), is 1 (but, given *k* alone, less than 1) [Dretske, 1981, p65]

Where *r*, the 'signal', is understood to be "any event, condition, or state of affairs the existence (occurrence) of which may depend on *s*'s being *F*" [Dretske, 1981, p65]. We may think of a neuron firing, or the position of a needle on a gauge. The condition '*s* being *F*' is to be understood as some item being in some state, or having some property, such as a chair being blue, or a petrol tank being empty. There must be some "positive information associated" with the condition, which requires that "there are possible alternatives to *s*'s being *F*" [Dretske, 1981, p65]. The chair may have been red, or the petrol tank may have been full.

By *k* Dretske means "what the receiver already knows (if anything) about the possibilities that exist at the source" [Dretske, 1981, p65]. As Dretske points out, there will be situations in which what you learn from a signal differs from what I learn. I might vaguely recall that a flashing red light on my phone means either it is low battery or the hard drive is failing. You know that a flashing red light means low battery and is unre-

lated to the hard drive. So, when you see the red light you learn something I do not - that the battery is dying. I only know that this is one of two possibilities.

Dretske points out that the knowledge condition, *k*, "constitutes a *relativization* of the information contained in the signal because *how much* information a signal contains, and hence *what* information it carries, depends on what the potential receiver already knows about the various possibilities that exist at the source" [Dretske, 1981, p79].

This relativization is consistent with classic information theory. James Mattingly provides a non-psychological interpretation of Dretske's theory in which he notes that, in information theory, calculating the amount of information which is transmitted to a receiver requires an understanding of "the decoding that is yet necessary for me to discover exactly what the content of the message is that I have received" [Mattingly, 2021, p192].

This may come as a surprise, since elsewhere Dretske famously maintains that information is "an objective commodity, a thing whose generation, transmission, and reception do not require or in any way presuppose interpretive processes" [Dretske, 1981, pvii]. However, Dretske is careful to distinguish information and informational *content*, which he considers as information which is, in some sense, available to the system itself. A representation may hold a wealth of information, but informational content is a subset of that information which can be interpreted by the system.

This consideration is what leads Dretske to assert that the conditional probability of the content on the representation be unity. Unity is necessary, argues Dretske, since it ensures that the "signal carries as much information about *s* as would be generated by *s*'s being *F*" [Dretske, 1981, p63]. Dretske considers this important because it allows the system to reduce the relevant alternate possibilities to none, enabling precise spec-

ification of content. If I want to know which side of a coin is facing up with absolute certainty, I must be able to reduce the possibilities by half - which requires precisely the same amount of information as generated by the source (i.e. 1 bit). Any less information would leave equivocation, or a form of uncertainty, about the external item.

This means that, for Dretske, the system must have as much information available to it about an item as is generated by the item itself for that item to be the content of the representation. The representation may, according to some objective measure, carry information about numerous other items, but unless those other items share a conditional probability of unity with the representation, they will not be part of the informational content of that representation.

In later parts of the thesis (for example, section 7.5) I will argue that Dretske is basically right. However, I will argue that a looser statistical measure, maximal mutual information, ensures the kind of availability Dretske is after. I will also argue (for example, in 6.5) that we need to add in functions. In the next section, I explore the work of Ruth Millikan, who, along with David Papineau, first drew philosophical attention to the importance of biological functions for a theory of content.

2.3 Ruth Millikan and functions

Ruth Millikan provided one of the first[7], and arguably most elaborately articulated, theories of representational content in terms of etiological functions. Despite apparent differences between Millikan's view and the informational teleosemantic framework, her position shares a number of key features. For example, as Millikan has emphasised in

[7]David Papineau [Papineau, 1984] advocated for a similar view around the same time as Millikan.

several places (see below), her view requires "correlations" and "mappings" between representations and contents, as with informational theories. Basic tenets of her approach also tend to be held by contemporary informational teleosemanticists (see section 2.6.1). As such, we can read Millikan as contributing significantly to the informational teleosemantic framework.

Millikan's view is given in *Language, Thought and Other Biological Categories* (LToBC) as:

> P is an indicative intentional icon of whatever it maps onto that must be mentioned in giving the most proximate Normal explanation for full proper performance of its interpreting device as adapted to the icon. [Millikan, 1987, p100]

The phrase 'indicative intentional icon' is typically glossed as 'representation' (endorsed by Millikan, see [Millikan, 2024, p55]). A representation is an item with a function in a wider system in virtue of what it indicates being "identified" and used by the system [Millikan, 1987, p13]. To first approximation, this means that the system is "guided" in its activities by the thing it represents. This is the content of the representation.

According to Millikan, the content of a representation is given by "the most proximate Normal explanation" of how the representation's "interpreting device" successfully performs a given task which relies on the representation for its success [Millikan, 1987, p33]. The 'most proximate Normal explanation' explains why a behaviour (e.g. of an organism which 'interprets' or 'consumes' the representation) was historically successful for ancestors of a system (e.g. historical members of an organism's species). Further,

the explanation must be the "*least detailed*" possible explanation [Millikan, 1987, p33] of such success.

A natural way to spell out 'least detailed' is as an explanation without any causally mediating properties being mentioned. This should deliver an explanation in which the *explanans* is most perspicaciously related to the *explanandum*. For example, a frog is successful in acquiring food because there is an external item which its representation maps onto: ⟨*frog food*⟩. That the representation maps to the item ⟨*frog food*⟩ provides an explanation of how ancestor frogs were able to survive by using that representation: using it allowed them to get hold of food for frogs. This leaves aside any details regarding *how* this process happens, such as the transduction of light intensities at the retinal interface.

Millikan adds this condition since she maintains that the representation must function "as a sign or representation *for the system itself*" [Millikan, 1989b, p284; emphasis in original]. Millikan writes that the representation must be "true" "*as the consumer reads the language*" [Millikan, 1989b, p286; emphasis in original], or that a producer "must be designed to speak whatever language the consumer(s) of its representations can understand; a representation consumer must be designed to understand whatever language the producer(s) of its representations speak" [Millikan, 2024, p55]. For example, frogs only understand frog food, not light intensities at retinal interfaces, so those proximal causal steps cannot feature as the representation's content.

In LToBC Millikan argues, not unlike Dretske, that this means that some possible content ascriptions are untenable because they suggest that the system itself has access to too much information: "VonFrisch knew what bee dances are about, but it is unlikely that bees do. Bees just react to bee dances appropriately" [Millikan, 1987, p13]. So, she

claims that the bee dance is not an instance of representation proper, since dances "do their jobs without their interpreters of the organisms that harbor them having any grasp of what they are about" [Millikan, 1987, p13].

In recent work, Millikan has provided her own psychologically neutral interpretation of representation for the system itself. She writes that "neural producers and neural consumers would both be substantiated in neural networks or whatever cognitive scientists and neurologists come up with next, involving multiple uses of parts of the nervous system depending on tasks to be performed" [Millikan, 2024, p55]. *Prima facie*, this overlaps significantly with the requirement that inputs be decodable by downstream neural systems which enact cognitive capacities (see section 7.5). Further, in *Neuroscience and Teleosemantics*, Millikan writes:

> A representation in the brain would have to be used as such by some other part or aspect of the brain or by something connected to the brain. If there are representations in the brain there must also be interpreters for them. [Millikan, 2021, p2460]

The 'interpreters' in question are downstream cortical areas which play some explanatory role in relevant cognitive capacities, such as actions an organism might do with an empty beer can: "stepping over it, picking it up with one's hands, picking it up with a stick, kicking it along the trail or into the woods [...] and so forth" [Millikan, 2021, p2461].

Millikan writes that for an organism to "'read' the code" of the representation, it is only "necessary that the creature should be guided by the signal in a way that diverts it from activities less likely to benefit it to ones more likely [to] benefit it" [Millikan, 2001,

p111]. Elsewhere, Millikan refers to the organism "understanding the language" of the representation in terms of what "'guides behaviour"' [Millikan, 2021, p2461].

In later chapters (for example, 3.3.2, 7.5), I will agree that content attribution requires that downstream systems be able to "read the language" of the representation, but that there are much more stringent requirements on reading the language, in an explanatorily salient sense, than merely being guided by the content. I will argue that downstream systems must be able to decode the content by performing mathematical operations over the values of the representation.

2.3.1 Functions

Millikan pioneered the use of functions in a theory of content. Millikan argues that functions enable *mis*representation. As Millikan spells out in multiple places, misrepresentation, on her view, is derivative of malfunctioning (e.g. [Millikan, 2021, p2465], [Millikan, 1990, p156], [Millikan, 2024, p56]).

Another reason Millikan introduces functions is that mapping relations alone are not sufficient to determine content, since "mathematical mapping relations are infinitely numerous and ubiquitous whereas representation-represented relations are not" [Millikan, 1987, p86]. Functions are thought to reduce the ubiquity of pure mapping relations by specifying those mapping relations which are phylo- or ontogenetically adaptive for an organism.

Millikan relies on an etiological notion of function, "proper function" [Millikan, 1987, p28]. Essentially, a system S has a component with a proper function F to do A, provided that ancestors of S who performed F and achieved A *thereby* successfully reproduced.

The heart has a proper function to pump blood, in this sense, because it was in virtue of the heart pumping blood in our ancestors that we were able to survive and reproduce (not, for example, because it made a whooshing sound - that was causally irrelevant to our survival, probably).

In chapter six I will argue extensively for the use of a non-etiological notion of function, while maintaining the use of functions broadly. I will also argue in this chapter (section 2.5) that etiological functions and non-etiological functions are two tools for different explanatory projects. The explanatory project I am pursuing differs from Millikan's, and requires non-etiological functions.

2.4 Karen Neander and response functions

Karen Neander's central contribution to informational teleosemantics is her detailed case for the existence of response functions. Like Millikan, Neander argues that teleosemantics requires a "mix" of both an input-based and output-based approach to content determination (e.g. [Neander, 2017b, p125]). Neander develops the input-based element of the account in terms of response functions, which are "functions to respond to something by doing something. Sensory-perceptual systems have functions to respond to various changes in the environment by changing their inner states in various ways" [Neander, 2017b, p126].

Neander elaborates on this definition by stipulating that to "respond to something (as I use the term 'respond' here) is to be caused by something to do something" [Neander, 2017b, p127]. For instance, "to say that a visual system changed into a RED-type state in re-

sponse to an encounter with a red visual target (due to its redness) is to say that the visual target's instantiating red caused (i.e., causally triggered) the system to change into a RED-state" [Neander, 2017b, p127]. Having a *function* to respond, meanwhile, means that the relevant system was "selected for responding to red being instanced by changing into RED-states" [Neander, 2017b, p127]. So, having a function to respond to something (x) by doing something (y) requires that a system *is* causally triggered to do y by an external item x, but also that it is *in virtue* of doing y in the presence of x that the system was selected.

This requirement is made explicit in Neander's "simple causal-information version of teleosemantics (CT)":

> CT: A sensory-perceptual representation, R, which is an (R-type) event in a sensory-perceptual system (S), has the content *there's C* if and only if S has the function to produce R-type events in response to C-type events (in virtue of their C-ness). [Neander, 2017b, p151]

In this formulation, 'there' "is used as a placeholder for the localization content of the representation" [Neander, 2017b, p152] - i.e. the location of the relevant external item (either relative to the visual field of the organism, or relative to a more distal measure, depending on the localization capacities of the organism - see [Neander, 2017b, p113]). A 'function' is to be understood in an etiological sense. However, as I will argue below (section 2.4.1), non-etiological functions better serve her explanatory aims.

The phrase 'in virtue of' is to be read in terms of what the system S was historically selected for. However, this itself should be understood causally; Neander argues that C must be "a causal difference-maker with respect to R-production by S" if it is to be selected

for [Neander, 2017b, p152]. If some non-C but C-like item would also produce R in S, then, Neander argues, C would not be the (sole) content of R. Rather, either R would have another content altogether, or R would be indeterminate with respect to C and the C-like item. This is because, Neander argues, an environmental item being a causal difference-maker enables specific capacities to be enacted when certain environmental conditions obtain. For example, "it was by responding to the dimming of light that the pineal gland produced sleepiness at nighttime" [Neander, 2017b, p132]. In this case, Neander argues that the pineal gland has the function of responding to dimming light.

2.4.1 Limitation to discriminatory capacities

According to Neander, the content of a representation must be constrained by the observed discriminatory capacities of the representation's sensory input systems. This is due to the fact that the response function of the representation requires that the external item must make some *causal* contribution to the activation of the representation. So, only those items which sensory systems are causally responsive to can stand as the content of a representation as determined by the response function of the relevant system.

The observed discriminatory capacities of the system are given by those environmental conditions which are experimentally determined to activate certain behavioural sequences [Neander, 2017b, p102]. Scientists discover which items induce the relevant behaviours, and which do not. To illustrate the point, Neander cites studies which investigate which external items engage a toad's prey-capturing behaviour. She reports that ganglion cells in the toad's visual system, the activation of which control its prey-capturing behaviour, are causally sensitive to the "location, size, shape, motion, and di-

rection of motion relative to the shape of a stimulus" but are "causally insensitive to whether the target is nutritious for the toad" [Neander, 2017b, p116]. The toad can represent size, shape, motion etc. but not anything like things which are nutritious for toads. So, content ascriptions such as those endorsed by Millikan - the toad is representing ⟨*toad food*⟩ - premised on the content being something nutritious for toads, are considered to be misguided.

However, Shea suggests a possible argument against limiting content to discriminatory capacities. Shea states that he does not "see why longarmed etiological functions need be tied to discriminative capacities" [Shea, 2018, p160]. Longarmed functions are so-called because they are determined by Normal explanations (e.g. [Shea, 2018, p159]), which provide the "*least detailed*" explanation [Millikan, 1987, p33] of evolutionary adaptiveness; they are 'longarmed' in the sense that they skip over intermediate explanatory stages, such as - crucially - the specific environmental conditions the system causally responds to. This is a problem, since Neander intends to invoke the very same notion of function. She intends the discussion to capture the kind of functions invoked by Millikan: "since the main parties to the present dispute agree on this [analysis of functions], I shall assume the etiological theory" [Neander, 2017b, p127]. However, she also wishes to maintain a restriction on content to what the system is causally responsive to.

Consider the pineal gland example; Neander writes that "it was by responding to the dimming of light that the pineal gland produced sleepiness at nighttime" [Neander, 2017b, p132]. However, strictly speaking, it was by responding to the dimming of light that the pineal gland produced sleepiness when the light dimmed. It was the fact that dimming light *correlated* with the onset of nighttime in the evolutionary history of the organism

that sleepiness was produced at nighttime. The function was selected because of this correlation - it reaches out, longarmed, to nighttime - not because of the way in which nighttime was tracked. There are always a number of logically possible ways to track an environmental item, but it is what is ultimately tracked which explains why any one of them was replicated. From an etiological perspective, the pineal gland has the function to - is there because it did - respond to *nighttime*.

This is brought out by applying the method of difference; had the item with which dimming of light correlated with been different, some other response would have been selected. If dimming light correlated with the onset of daytime (followed by a sudden brightening), the pineal gland would not have been selected for inducing sleepiness. If the onset of nighttime had been accompanied by a sudden bright flash followed by immediate darkness, the pineal gland would not have been selected for at all (at least, not in this capacity), or it would have responded to a bright flash followed by darkness. Etiological functions are not limited to discriminatory capacities; only the explanation for *how* the adapative environmental item was tracked *is* so limited. This, I will argue, is the role of C-functions - they explain how some cognitive capacity is enacted. Etiological functions do not.

So, Shea rightly points out that etiological functions do not need to be tied to discriminatory capacities, but Neander *requires* functionally-determined content to be limited to discriminatory capacities. This is because Neander argues both that content attribution must be constrained by scientific practice and that the relevant science limits content attribution to discriminatory capacities (e.g. [Neander, 2017b, p137]). So, it is a serious problem for Neander if the above argument is right, and etiological response functions

do not need to be tied to discriminatory capacities. Something has gone wrong. In the next section, I argue that etiological functions are not fit for the type of how-question explanations given in the regions of cognitive science which Neander is interested in.

2.5 Different explanatory projects

In this section I argue that implicit in teleosemantics are two different but complimentary explanatory projects. There are those which deal with how-questions and those which deal with why-questions. Some projects aim to explain *how* a given cognitive capacity is enacted. Other projects aim to explain *why* some representational state is present.

The distinction between how- and why-questions is due to Mayr in the context of his discussion of the difference between functional biology and evolutionary biology. According Mayr, the functional biologist seeks to answer the question: "How does something operate, how does it function?" whereas the evolutionary biologist seeks to answer "the historical 'how come?"' [Mayr, 1961, p5102] - the historical reasons some item is the way it is. Mayr refers to answers to how-questions as "proximal" explanations and answers to why-questions as "ultimate" explanations.

As Neander states, "'How-questions' and 'Why-questions' are mnemonic tags that name questions that are often but not invariably asked by using the words 'how' and 'why."' [Neander, 2017b, p256]. With some linguistic ingenuity we can translate almost any question from a 'how' to a 'why' or vice-versa. The relevant distinction is that an ultimate explanation goes beyond what we can infer from experimental conditions and has to do with the historical conditions of the system or its ancestors.

How-questions, aimed at addressing proximal explanations, are concerned with what the organism itself is able to do, including any limitations it faces. Mayr spells this out in the case of functional versus evolutionary biology:

> We can use the language of information theory to attempt still another characterization of these two fields of biology. The functional biologist deals with all aspects of the decoding of the programmed information contained in the DNA code of the fertilized zygote. The evolutionary biologist, on the other hand, is interested in the history of these codes of information and in the laws that control the changes of these codes from generation to generation. [Mayr, 1961, p5102]

There is a deep analogy between what Mayr is describing and the two projects as I conceive them. The philosopher or scientist pursuing how-questions *should*, I will argue throughout this thesis, be concerned precisely with what the system itself can decode from its input, in a strictly information-theoretic sense. In contrast, the philosopher or scientist pursuing why-questions can be construed as asking why it is that some environmental item is encoded and decoded, which in turn will explain the evolution of the cognitive system over evolutionary (and learning) history.

Confusingly, both the how-question and the why-questions are partly answered by way of a what-question. In the case of the how-question, the what-question is "what does the system itself pick up (encode) and use (decode) from the environment in order to perform its cognitive function?" Studies which attempt to find answers to this question are the type explored in the next chapter, where we will look at cognitive neuroscientific

accounts of face recognition. Such studies attempt to discover precisely which aspects of the environment are processed by the cognitive system.

Why-questions invoke another type of what-question: "what, in the environment, correlated with what is encoded and decoded, is adaptive for the organism?" This involves looking beyond the processing limitations of the system itself to those environmental elements which *we* know are beneficial to that organism, but about which the organism itself may have little to no conception. Another way of formulating the why-question would therefore be something like "why does the system encode and decode what it does?" For example, why does the system encode and decode dimming light? Because dimming light correlates with nighttime, and it was adaptive for the system to track nighttime (by way of the dimming light).

This can lead to some apparent disagreements. We might say that the face recognition system represents $\langle$*two dots above a line*$\rangle$, or we might say it represents $\langle$*face*$\rangle$. In other words, we might advocate for what is known as a "low-church" reading of content, or we might advocate for a "high-church" reading. Karl Bergman clarifies the two concepts:

> We can think of low contents as hewing closer to the perceptible features that are directly involved in causing the representation to be tokened, whereas high contents are concerned with the ecologically relevant features that are involved in explaining the evolutionary success of behavior guided by the representation [Bergman, 2021].

I claim that whether we take the content of the relevant representation in the facial recognition to be either high or low-church depends on our explanatory project. If we want to explain *how* the system performs face recognition, we should go low-church:

this is what the system itself encodes and decodes from the environment. If we want to explain *why* the system has this representational state we should go high-church, identifying the item which the low-church content is correlated which was adaptive for the organism. Cognitive science is a diverse discipline, encompassing a number of approaches. A full understanding of any cognitive system will involve answering both how- and why-questions.

This suggests that there is no serious disagreement between teleosemanticists, just a difference of explanatory project. Rather than consider there to be a disagreement over what the content of a representation is, we should think of 'content' as a word which describes two broadly similar but distinct environmental items. Both items have some relation to an organism, and both items may be explanatory with respect to the same cognitive capacity. However, *which* relation they have, and *which* aspect of the cognitive capacity they explain is different. We are talking about different things, because we are pursuing different projects. Maybe one of us should stop using the word 'content' - but old habits die hard.

How-questions and C-functions

Neander's project is aimed at answering how-questions. She contends that "citing the normal-proper functions [for Neander, etiological functions] of the components of a system can play a significant scientific role in the answers to How-questions that physiologists and neurophysiologists provide" [Neander, 2017b, p60]. Given that Neander is attempting to provide an account of content consistent with cognitive scientific answers to how-questions, she aims to be consistent with that region of cognitive science which

deals with those how-questions. For example, she characterises the question guiding the study of the toad, which she uses to motivate her position, as: "How does the motivated toad distinguish preylike from predator-like and other moving stimuli?" [Neander, 2017b, p105].

Like Neander, I think that scientific answers to how-questions involve the positing of content. Unlike Neander, I do not think etiological functions are apt for answering how-questions. As mentioned above, this is because etiological functions are not tied to the discriminatory capacities of organisms. So, etiological functions are inconsistent with the use of experiments to test the discriminatory capacities of organisms, performed in order to assess how an organism performs some behaviour.

However, the non-etiological functions identified by Cummins (what I call 'C-functions') *are* tied to the processing limitations of organisms. They are restricted to what a system can occurrently *do*, not with what its ancestors were adapted to. This makes my account in line with the generally-accepted position, as articulated for example by Griffiths [Griffiths, 2006, p3] and Millikan [Millikan, 1989a, p175], that C-functions provide answers to how-questions. Cummins defines functions as:

> x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s's capacity to ψ, just in case x's are capable of ϕ-ing in s and A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to ϕ in s

For example, an we may want to know whether the system itself is capable of responding to nighttime, or whether it is only capable of responding to dimming light. An 'analytical account' of how the system does this requires finding what the system

actually, here and now, responds to. In turn, this involves analysing the system into subsystems, and identifying how each subsystem performs under a variety of experimental conditions. Using this method, we will discover that what the pineal gland actually *does* is respond to dimming light across a range of conditions. If the hypothesis were that the pineal gland, here and now, responds to nighttime, we would struggle to explain experimental data indicating that the pineal gland also responds to a range of *other* conditions, and under a number of conditions fails to respond to nighttime. In other words, this hypothesis accounts for far less of the evidence than the hypothesis that it responds to dimming light. However, on an etiological account this is a good hypothesis: if we look at the evolutionary history of the species, we will find that the activation of the pineal gland was adaptive when it occurred at nighttime.

In summary, C-functions provide an account of what a subsystem occurently does within a wider system, which is discovered by using experimental conditions and observing what the target component does in those conditions, then providing the best explanation of its activity. Since what the component does is limited by its discriminatory capacities, our best explanation will track those discriminatory capacities, and the ascribed function will be limited to those discriminatory capacities.

Relation between projects

The two projects are compatible and mutually supportive. High-church content ascriptions can help researchers working on how-questions narrow the range of possible low-church contents by delimiting a region of items that the system is likely responsive to, if it is to fulfill its etiological function. For example, if we have a hypothesis which states

that it is adaptive for humans to respond preferentially to faces, we might look for regions of the cortex which activate when presented with faces. Once we find these regions, we can increase the specificity of our search, successively narrowing the experimental conditions and manipulating the environmental variables until we find what that region is preferentially responsive to. This will allow us to isolate the particular element of the environment which is the low-church content.

Low-church contents can help answer why-questions. As mentioned above, we can apply a why-question *to* the existence of low-church content. We might ask: why does that region of the facial recognition system respond to shape properties? We might answer that in the evolutionary history of the organism, such properties tended to coincide-With faces, allowing successful recognition.

If we use the kind of studies outlined throughout this thesis to discover low-church content, we can open up a whole realm of theorising in the domain of why-questions. We can essentially provide new input data to those theories. If we begin with a why-question, pursuing the appropriate methodology, we can then apply our how-questions and find the specific environmental items that the organism uses to fulfill its etiological functions and thereby discover precisely how it works.

A comprehensive understanding of anything requires a number of researchers pursuing various types of questions which can be asked of that thing. Understanding cognition is no different. We can ask both how cognitive capacities work, and why they have been historically adaptive. An answer to each question enriches our understanding of the cognitive system.

2.5.1 Section summary

Teleosemanticists working on content determination pursue different projects, and use corresponding regions of cognitive science to support their account. Those working on etiological theories answer why-questions concerning the existence of a given mental representation. Those working on non-etiological, often causal or informational, theories answer how-questions concerning the occurrent operation of the system. Neander argues that etiological functions can answer how-questions. They do this by way of supporting response functions. I agree that response functions are crucial, but I have argued that etiological functions are not limited to discriminative capacities. The relevant regions of cognitive science *do* posit contents which are limited to discriminative capacities. So, etiological functions cannot be used to answer how-questions (if they are to be consistent with scientific answers to these questions, which Neander intends them to be).

Some confusion has been generated by the fact that both projects require answers to what-questions; they each posit some external environmental item which has some explanatory relation to the cognitive capacity under investigation. The nature of this explanatory relation, given the difference in explanatory project, is different. The items which are posited as content are also different. However, this does not mean the projects are in competition. Rather, they inform and support one another to generate a full understanding of the cognitive system.

2.6 Contemporary teleosemantics

Theorists do not always specify the explanatory project they are pursuing. However, I have split the accounts into those which I argue are best interpreted as aimed at either why- or how- questions. This should help clarify the state of the art relative to each type of project.

2.6.1 Answers to why-questions

Peter Schulte

Schulte provides an account built largely on Neander's account of response functions. However, Schulte also incorporates constancy mechanisms as a crucial component of the account which, he argues, allows the producer-based teleosemanticist to overcome the "distality problem". The problem is that it appears as though producer-based teleosemantic accounts struggle to isolate non-proximal contents, such as the immediate pattern of retinal firing giving rise to ganglion cell responses.

Schulte follows Neander's example of the toad's visual system, in which Neander argues that the +T5(2) neuron represents "small, elongated objects which move in the direction that parallels their longest axis" which Schulte calls "SEM objects" [Schulte, 2018, p353].

The distality problem applied to this example produces the following question: why "is it the case that +T5(2) represents the presence of a SEM object and not the presence of certain retinal stimulation pattern, or a certain pattern of light?" [Schulte, 2018, p355]. Each proximal item is as much of a cause of the relevant neuron firing as the more distal

item. Since Neander's account of response functions is a causal notion, it looks like the cell's response is indeterminate between these causes.

Neander's solution to this problem involves pointing out that "pathways in the toad's visual system were selected for responding to the light by producing certain tectal firings because by that means they responded to the distal worm-like motion, and not vice versa" [Neander, 2013, p35]. Neander invokes an asymmetrical in-order-to relation in order to deliver distal content. The content is the highest point of the in-order-to relation: each proximal causal interaction is present because it allows the system to respond to the more distal item, while the distal item is *not* responded to in order that the system can respond to the more proximal items.

Schulte points out a problem with this response. Invoking another example, he suggests that potassium-richness is causally relevant for producing [neural response] T, since it is causally relevant for producing the insects' red surface colour; so Neander cannot deny that the toad's visual system has the function of producing T in response to potassium-rich objects" [Schulte, 2018, p359]. This is "in conflict with Neander's view that (basic) perceptual states represent the surface features of objects" [Schulte, 2018, p359]. It looks like Neander's solution to the distality problem produces content ascriptions which are *too* distal.

Instead, Schulte suggests that we turn to "constancy mechanisms" to provide the appropriate level of distality. For a detailed description of constancy mechanisms (sometimes also called "invariance mechanisms") see chapter five section 5.4.1. In short, constancy mechanisms allow the system to respond to the same external item across a range of proximal conditions. Imagine a toy system with a 'red ball' neuron connected to some

eyes. The 'red ball' neuron fires whenever there is a red ball present and visible to the eyes. However, the red ball may project light onto very many areas of the retina. The 'red ball' neuron fires regardless of where on the retina the light is projected: its response is invariant across a multitude of proximal retinal conditions.

Applied to the toad example, Schulte observes that "there is no single type of retinal stimulation pattern which normally causes the toad's visual system to produce T5(2) activation; instead, T5(2) activation is produced in response to very different retinal stimulation patterns under different circumstances. The same seems to hold for patterns of light. Hence, the only external state that qualifies as a normal cause of T5(2) excitation, i.e. as a cause that is always present in normal situations, is the distal state [a SEM object is present]" [Schulte, 2018, p361]. Schulte argues that the represented item is that item which provokes the invariant response across each proximal condition.

However, Schulte notes a possible objection: "appearances to the contrary, there *is* a proximal state that qualifies as a normal cause of T5(2) activation — namely, a disjunctive proximal state" [Schulte, 2018, p362]. If we could create a list of all the proximal states which activate the relevant neuron, we could, it seems, equally well ascribe this disjunctive proximal set as the content of the corresponding neural representation.

Schulte suggests the following solution: "we can solve the distality problem by identifying the content of a perceptual state with its most natural (least disjunctive) normal cause" [Schulte, 2018, p363]. In this case, the most natural cause, [p], is more 'natural' than the disjunctive cause, [p*], just in case there is "a higher degree of objective similarity between [p]-tokens than there is between [p*]-tokens" [Schulte, 2018, p364].

Setting aside a detailed investigation of this notion, I wish only to remark that the

reliance on the more 'natural' cause renders Schulte's view more suited to answering why-questions than how-questions, at least as the theory stands. We have no indication of *how* the system is able to respond to 'natural' rather than 'non-natural' items. Instead, positing that the system *does* respond to 'natural' items can be used to explain *why* the system tends to be successful: if the 'natural' item is present, the system will be able to access whatever the item provides (e.g. nutrition). If a member of the disjunctive set of retinal activation patterns is present only, the system may nonetheless not be in a position to acquire nutrition (e.g. due to random activation of the same neural pattern) - hence the system will fail to gain sustenance.

Schulte's invocation of constancy, on the other hand, provides an invaluable contribution to informational teleosemantic accounts aimed at determining contents which can be used to answer how-questions. I go into detail about how these mechanisms allow the system to define an external random variable over which mutual information quantities can be calculated in chapter five. In chapter seven, I will suggest that, rather than invoke 'naturalness' to solve Schulte's disjunction concern, we can instead defer to what the system itself can *use* (decode) to perform its tasks.

Manolo Martinez and Marc Artiga

Martinez and Artiga each provide accounts which make an explicit attempt to provide ultimate explanations. High-church content is taken to be explanatory of the evolutionary persistence of so-called "indicator" states. As Artiga puts it, content ascriptions on these accounts are intended to "explain why the representational system was selected for" [Artiga, 2021, p473]. I quote an example of his in full:

> Dragonflies possess a set of neurons called "target-selective descending neurons" (TSDN), which are completely silent unless the dragonfly is presented with a target within the adequate receptive field, with a certain size (about 1–2°) and moving in a determined direction (Olberg, 2012; Sathe & Bhusnar, 2010). Activity in TSDN causes dragonflies to quickly move in a certain trajectory, which in an astonishing 95% of cases allows them to catch prey (Combes, Salcedo, Pandit & Iwasaki, 2013; Gonzalez-Bellido, Peng, Yang, Georgopoulos & Olberg, 2013, p. 699; Olberg, Worthington & Venator, 2000, p. 155). If for the time being we make the simplifying assumption that dragonflies only prey on mosquitoes [...] teleosemantics would entail that activation in TSDN is a representation of something like a mosquito being around [Artiga, 2021, p472].

The content ascription, roughly $\langle mosquito \rangle$, is made on the basis that it explains *why* a dragonfly has TSDN neurons. Generally, such contents are intended to answer why-questions of this sort. The content ascription also explains *why* the dragonfly is successful as often as it is - for example, perhaps the low-church contents of TSDN neurons are very highly correlated with mosquitoes (for example, TSDN neurons may have as low-church content a highly specific shape profile). As Martinez writes: "the simplest contentful states do exploit (and exist because they exploit) a correlation between detectable (low) properties (Being shiny and black, say) and useful (high) properties (e.g., Being nutritious, or Being dangerous)" [Martinez, 2013, p441]. However, I argue that if we want to know *how* the dragonfly performs this cognitive feat, we must have some theory which can determine precisely what its TSDN neurons represent in the 'low' sense.

Martinez also considers etiological content-ascriptions to be aimed at answering why-questions: "To provide a content attribution for a representation type R is to provide a compressed explanation of the existence of R in a sufficient number of cases." [Martinez, 2013, p450]. Martinez also considers any etiological-function-based theory of content to be designed to address why-questions: "at least under the etiological understanding of what functions are, to say that something is whatever a representation has the function to indicate (i.e., its content) is to say that it figures in an important way in an explanation of the existence of the representation" [Martinez, 2013, p451]. This is to say it answers *why* the representation is present - or *why* the system discriminates what it does. As I use the terminology, this type of high-church content is used to answer *why* the representation has the (low-church) content the science attributes to it - which it attributes in order to answer how-questions.

The question both Martinez and Artiga set out to answer is how to make a principled attribution of an item, correlated with the low-church content, as the high-church content of a representation. Given that many environmental items are correlated with the low-church content the system has, which item should be selected?

According to Martinez, "something similar to Boyd's HPCs solves the indeterminacy problem for naturalistic accounts of content" [Martinez, 2013, p443]. HPCs are 'homeostatic property clusters', which "are individuated by property clusters that are afforded imperfect yet homeostatic integrity by underlying causal mechanisms" [Wilson et al., 2007, p190]. Martinez also stipulates that "the clustering must be causally important for the existence of the indicator M" [Martinez, 2013, p444]. The correlated item is given by the cluster of properties that what is represented (in the low-church sense) is causally related

to, by way of mechanisms which support the homeostatic integrity of the cluster.

Artiga argues that Martinez's account fails, since we can in principle discover a number of nested HPCs. A wing, for example, may be an HPC made up of all the properties of a wing which allow it to keep its integrity. The wing may also be causally related to the low-church content of representations such as those enacted by TSDN neurons, making Martinez's account indeterminate between the contents $\langle mosquito \rangle$ and $\langle wing \rangle$.

Artiga proposes that "the [high-church] content of a given representation is the property F_1 that best explains why the other properties in P tend to co-occur" [Artiga, 2021, p481]. Artiga provides an example: "the presence of a mosquito involves a collection of well-established mechanisms that strongly tends toward the production of wings and the subsequent ability of fly. In contrast, note that no mechanism has been found leading from wings to the production of mosquitoes" [Artiga, 2021, p484]. In this case, *being a mosquito* is the property F_1, and this explains why wings are there, and why specific low-church properties are represented by the system. The reverse does not hold. So, Artiga's account is argued to produce the content $\langle mosquito \rangle$ for TSDN neurons.

Both approaches look promising for answering why-questions. However, they also demonstrate the need for a systematic account of low-church content. We need to be able to provide a principled specification of the precise environmental items that the TSDN is representing in our sense if we are to discover the best candidate property cluster - either homeostatic or explanatory. To know which set of items to look for, we need to be able to identify precisely some of the items *in* that set - the environmental items picked up by the system itself. I believe that maxMI can provide just such an account.

2.6.2 Answers to how-questions

Typically, answers to how-questions are sought by those interested in *structural* representation. This is the area of research most aligned with the project of this thesis. In order to illustrate the approach, I will highlight a significant contributor to this research area, Gualtiero Piccinini.

Piccinini defines structural representations roughly as "a model of a target that can guide behavior with respect to its target" [Piccinini, 2022, p4]. Piccinini glosses this using the analogy of a map: most maps represent their 'target' (the terrain) by being structurally similar to the target; spatial properties of the environment are reproduced on the map: if the river is to the left of the hill, a map will typically show a diagram of a river to the left of a diagram of the hill. In technical terms, structure-preserving maps are "homomorphic to the systems they represent" [Morgan and Piccinini, 2018, p15]. A change in an external item, given a homomorphism, leads to a change in a representing system.

Importantly for structural theorists, content features in an explanation of *how* the system's behaviour is guided. As such, it must be the case that "the information content of those states is explanatorily relevant to what the system does" [Morgan and Piccinini, 2018, p15].

How is the explanatorily relevant content determined? According to Piccinini, the best account of "the semantic content of structural representations is informational teleosemantics, which says, roughly, that the semantic content of a structural representation is the information it has the function of carrying about its target" [Piccinini, 2022, p5]. This is surprising, since structural representation is often thought to be at odds with teleosemantics. However, Piccinini notes that "it's a basic corollary of communication theory

that if one system encodes information about another, a homomorphism holds between them. Thus anything that qualifies as a receptor ipso facto qualifies as a structural representation, and vice versa" [Morgan and Piccinini, 2018, p15]. So, informational teleosemantics - at least a version which uses information theory proper - aligns with work in structural representation.

In order to see how content is explanatorily relevant for how-questions, Piccinini suggests that we must find an account "that sheds light on explanatory practice in the cognitive sciences, notably cognitive neuroscience" [Morgan and Piccinini, 2018, p15]. We can find such an account by observing the fact that the cognitive neuroscientists' methodology involves "investigating the response properties of neurons and neuronal populations", allowing them to "determine what such neurons or populations are most responsive to under relatively good sensory conditions". According to Piccinini, "that is their semantic content" [Piccinini, 2022, p10]. This means that "the content of individual neural representations is for neuroscientists to investigate empirically, not for philosophers to intuit about" [Piccinini, 2022, p10]. In this respect, the current project is in complete agreement.

An independent argument for Piccinini's position, that cognitive neuroscience is the place to look for content featuring in explanations for how-questions, is provided in chapter five section 5.5. In short, taking information theory seriously requires finding brain structures which can be modelled as mathematical entities called 'random variables'. Psychologically-characterised mental states cannot provide the requisite precision in terms of identifiable outcome values and the probabilities of those values which are used for random variable modelling.

However, as we will see over the course of this thesis, content determination in cognitive neuroscience is more complex than it initially appears. In order to be truly explanatory, content must be *decoded* by downstream systems. Simply looking at the item that "neurons or populations are most responsive to" does not guarantee that the *system itself* can decode the same information which *we*, with all our experimental background knowledge, can decode. This is why a significant proportion of cognitive neuroscientists are now explicitly attempting to use empirical methods to discover the information which is decoded, rather than information which is available from the neuron's response profile for the experimenter.

Cognitive neuroscientists do occasionally look for the item that the neuron is "most responsive to". For example, classic studies on visual object recognition such as those by Tanaka and colleagues (e.g. [Tanaka, 1992], [Tanaka, 1997]). However, many contemporary cognitive neuroscientists use more sophisticated measures, such as the spike-triggered average (STA) of the cell, dimensionality-reduction, or conditional mutual information. When we look at these methods, we see that they implicitly rely on the content of a neural representation being that item with which the representation maximises mutual information. The use of functions and the implicit commitment to maximal mutual information suggests that the implicit theory of content operative in cognitive neuroscience is maxMI.

2.6.3 Information theory

Cognitive neuroscience, as I argue in chapter 4 section 4.4, relies on Shannon's information theory - either implicitly or explicitly - in order to characterise the external item

which is explanatorily relevant to how-questions. However, many theorists[8] contend that Shannon's information theory is an inappropriate tool for a theory of content. They argue that information theory is not concerned with the *content* of signals, merely with the *quantity* of information contained within them. So, it cannot be used to deliver content. However, recent work attempts to show that information theory *is* an appropriate tool for a theory of content.

Stephen Mann

Stephen Mann answers the above challenge head-on by arguing that Shannon has been misinterpreted in his famous 'warning' that information theory has nothing to do with content. Mann writes: "I demonstrate that Shannon's warning pertains to sources, not signals. When Shannon did turn to signals, he called them representations and explicitly referred to them as contentful" [Mann, 2023, p9]. Shannon's warning, in this context, is:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at an- other point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. *These semantic aspects of communication are irrelevant to the engineering problem* [Shannon, 1948, p379; emphasis added].

Mann argues that philosophers have routinely misread Shannon's warning. He contends that Shannon was referring to the *source* message rather than the *code* string. Whether the source string is meaningful or nonsense does not matter for how much

[8] e.g. [Dretske, 1981, p41], [Neander, 2017b, p7], [Lombardi et al., 2016, p1989].

information is available from the source - this just relies on there being the some number of possible states with some probability distribution over each state. However, this does not mean that measures such as *mutual information* do not involve content. As Mann writes, Shannon's warning does not pertain to "a correlation between signals and signifieds" [Mann, 2023, p13].

The positive case that mutual information, as a measure, *does* import content is given by Mann's demonstration that information theorists themselves implicitly assume that coded messages are contentful in their theoretical work. He argues that Shannon's 'source coding theorem' (a theorem about how many symbols are required to encode a source) "assume[s] that the code symbols are being used to record the symbols of the source string" [Mann, 2023, p14]. Absent the assumption that the symbols received at the receiver are 'about' the source, the source coding theorem makes little sense. To determine how efficiently a source has been encoded, we need to be able to compare the code string to the source string - which it is *about*.

While Mann does not attempt to show how we can use information theory in theories of content, he does clear a significant hurdle: he shows that "the idea that formal tools are conceptually and technically distinct from the philosophical question of semantic content is becoming untenable" [Mann, 2023, p23]. Mutual information, in particular, is a notion which can in principle be deployed in search of a theory of content. Nothing in the formal model prohibits this. As I will show throughout the thesis, but especially in chapter seven (section 7.4), we can use amount of mutual information to provide a principle for content determination: maximal mutual information provides an item about which the system has 'available' information, isolating the item which provides explanations for

how-questions.

Oliver Lean

Lean aims to show that information theory is an appropriate tool to answer why-questions. He argues that "selection would produce organisms that respond adaptively to causes that carry mutual information about selectively relevant states of the environment" [Lean, 2014, p401]. When considering why some system within an organism is selected for, we need to consider the amount of mutual information it provided, historically, about the input. The more information carried by a representational structure about some item beneficial to the organism the more selection pressure there will be on that structure.

While information relations in Shannon's sense are ubiquitous, Lean argues that there is a role for Shannon information in "distinguishing processes on the basis of the quantity and fitness value of the information being discussed" [Lean, 2014, p404]. This limits the range of cases in which Shannon information provides a guide to content attribution. It is only when a system is *selected for* in virtue of its Shannon information that the system can be said to have the 'function' to deliver that information. Once this function has been established, we are able to determine *mis*representation. So, Lean argues, "Shannon information, in concert with function and fitness, can do the same explanatory work as richer informational concepts" [Lean, 2014, p406]. Specifically, it allows us to limit representation to explanatorily relevant items, and allows us to see how misrepresentation is possible.

Rosa Cao

Mann and Lean attempt to show that information theory, *pace* significant philosophical pressure to the contrary, is an appropriate tool for a theory of content. Rosa Cao considers how information theory can be applied to the cognitive system under a broadly informational teleosemantic approach. I will focus on her discussion of what it takes for some part of the cognitive system to be a 'receiver' of information. Specifically, Cao is interested in the properties a potential receiver must have in order to receive information from the external world

Referring to structural accounts such as those offered by Piccinini, Cao writes: "Some causal accounts of representation require that the signal correlate with or show isomorphism to a state of the world. I would argue that correlation by itself is not sufficient to give a signal its content" [Cao, 2012, p53]. As I alluded to above (section 2.5), what downstream systems *do* with the incoming information is crucial to determining representational content. Cao argues for this position and therefore focuses on the information which reaches the receiver. However, Cao notes that there is some ambiguity in attempting to characterise what counts as a receiver within the brain. Cao states the problem as follows:

> Are receivers pre-synaptic terminals or post-synaptic densities? Are they whole post-synaptic cells, or groups of neurons defined by common inputs? What about groups of neurons defined by projections onto common targets, groups of neurons defined by functional role, or groups of neurons defined by cell type? The grouping can also be vertical, in the sense that maybe the right way to think of a receiver in the brain is as the post-synaptic density—and

> everything it's attached to and has any effect on further downstream, all the way out to the neuromuscular junction and motor actuators. But then, downstream itself is a problematic concept in a brain full of feedback connections. [Cao, 2012, p60]

Cao highlights the fact that only some downstream systems can be said to truly receive information in the sense that they can use it. Many of the possible receiver systems she highlights are unable to use the information they receive as she conceives of 'use'. She argues that receivers must meet the requirement that they can "act" on the input. Her characterisation of this requirement is that anything which is a receiver must be an "agent", which she spells out as follows:

> the receiver needs to be an agent, again, not necessarily in the sense of having intentions, but at least in the sense of being able to act in the world to affect its own outcomes. An incoming signal will only carry semantic information for its receiver if the receiver has the ability to act on the information in a consequential way, even if that ability is not exercised often, or even perhaps has never been exercised. [Cao, 2012, p53]

Cao eventually argues that the only true receiver is the *whole organism* - that entity which is able to act on the environment. So, any sub-component of the organism, taken by itself, is unable to act on the environment in the requisite way. According to Cao, this means that we must "give up the common interpretation that a single neuron is doing something like representing 'a very small piece of the world outside the organism,' though indeed its activity may be well correlated with the structure of that small piece of the world" [Cao, 2012, p66].

Cao argues for this position due to familiar considerations about the fact that content must contribute to an explanation of the persistence of the representational structure (i.e. answers why-questions of the sort we explored above). This means that the putative receiver must be able to act on the item which is the content of its representational states, and receive a reward for doing so, thereby resulting in the stabilisation of the representation - its persistence over time. Any receiver which can thereby act on the world and receive a reward has access to the relevant information. However, single cells have no such access. She provides an example: "the V1 cell has no access to contour edges. What it does have access to is whether the cells that project to it are firing, and how much inhibitory vs. excitatory transmitter it is experiencing" [Cao, 2012, p66]. The V1 cell cannot act on contour edges or receive a reward from them. All it does is reliably respond to such edges, and transmit that information by way of 'acting' on its proximal input.

In essence, Cao is emphasising the role of the consumer for informational theories. According to her view, the content of any representation is given by its ultimate relation to some cognitive capacity. However, there is an alternate way to think about the content of sub-systems within the whole system, given this requirement. Some sub-systems contribute very specific environmental information to the operation of the whole capacity. For example, some cognitive tasks require recognising an object within a larger picture. When cognitive neuroscientists investigate this capacity, they find that some very specific elements of the picture contribute to the recognition of the whole object (e.g. [Zhan et al., 2019] - investigated in depth in chapter four section 4.4.3).

If we drop the requirement that content must feature in an answer to a why-question,

thus removing the need for a reward, we can attribute specific contents to subsystems in virtue of what they can be experimentally discovered to contribute to the whole capacity. On such a picture, what a receiver has access to will be spelled out in terms of what it can *decode* and transmit to other sub-systems (as spelled out in detail in chapter seven section 7.5).

2.7 Conclusion

Drestke introduced information theory to the content naturalisation project, providing fecund ground for a research project aimed at understanding representational content in terms of informational relations. Ruth Millikan emphasised the importance not only of informational relations, but "usable" information relations - limiting content to items serving some function of the system. Karen Neander further limited content by introducing response functions, in which content is tied to discriminatory capacities. However, perhaps Neander's most important contribution is noting the divergence of explanatory projects within teleosemantics, arguing that content can contribute to how-question answers, not just the why-question answers sought by many.

Contemporary researchers within the teleosemantic tradition have built on this foundational work in a number of ways. Schulte, Martinez, and Artiga each attempt to provide a way to systematically determine the item in the environment, correlated with what the system can discriminate, which is explanatory relevant to the why-question project. Those who answer how-questions, such as Piccinini, emphasise the role of multi-level componential explanations in cognitive neuroscience as a guide for content determina-

tion. Piccinini recommends observing how neuroscientists measure response profiles of neurons in order to determine representational content. While this is true, it undersells the complexity of the methodologies and implicit theorising which support content attributions in contemporary cognitive neuroscience.

Picking up from Drestke's employment of information theory, contemporary researchers attempt to answer outstanding questions about the suitability of information-theoretic methods for theories of content. Mann answers the criticism that information theory is irrelevant for the purposes of determining content. He suggests that information theory *is* concerned with content in the sense that signals *encode* a message, where this relation is spelled out in representational terms within information theory proper. Lean then argues that information theory can be used to answer why-questions, demonstrating the power of the model.

Cao investigates the applicability of information theory to the cognitive system, attempting to isolate mechanisms which can count as receivers within the information-theoretic sender-receiver model. She argues that receivers must be able to act on their content directly. However, another possible model involves discovering what a receiver is able to decode from its input in order to contribute to a wider cognitive capacity.

Generally, theorists acknowledge that structural representations play an important part in answers to both how- and why-questions. However, no theorist other than Neander has yet provided a systematic treatment of how content is ascribed to such representations. To speculate a little, perhaps theorists do not consider such content ascriptions theoretically interesting, since they appear to be able to be given by a relatively simple correlational relation. As Ramsey says, "functional theories of content typically presup-

pose that neural states function as representations; they do not explain how this happens" [Ramsey, 2016, p6]. In this thesis, I attempt to spell out precisely how content attributions - the type which are geared towards answering how-questions - are made-Within cognitive neuroscience. Rather than mere correlation, content ascriptions are highly constrained. As I argue, they must meet the requirements of maxMI.

Chapter 3

Content Naturalism

3.1 Introduction

Neander and Piccini, along with most contemporary informational teleosemanticists, turn to scientific practice to discover an implicit theory of content. This is also the methodology I pursue in this thesis. However, the philosopher Frances Egan argues that content ascriptions in cognitive science are determined by the pragmatic interests of researchers, and that content is not strictly part of scientific explanation. Content attributions, Egan argues, only serve various heuristic purposes, such as making the scientific theory perspicaciously related to the presumed pre-theoretic interests of readers. In this sense, contents are not naturalistic - they are determined only by pragmatic considerations. If so, there is no implicit theory of content in cognitive science which isolates contents which are explanatory of cognitive capacities, contrary to what we have been assuming so far.

Egan offers two considerations in favour of her position. First, that content is not "essential" and second that content is not "naturalistic". Egan argues that content is not essential in the sense that it is not used to individuate representations in cognitive science. 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 [Chang and Tsao, 2017], representational states are individuated with respect to their contents precisely in the way required by Egan (section 3.3.1). I then show that content is determined by the naturalistic relation of encoding, which constrains content attributions not with respect to the heuristic requirements of researchers, but by the discriminatory and decoding capacities of the system itself (sections 3.3.2, 3.3.3). I then argue that such content attributions are sufficiently determinate so as to count as explanatorily relevant contents (section 3.4).

I end by considering three principles which 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 3.5).

3.2 The problem

Many contemporary philosophers who attempt to provide a naturalistic theory of representational content (i.e. a theory which specifies how the content of any given representation is determined consistent with scientific practice) employ a methodology, expressed by William Ramsey as follows: "to critically examine the different ways cognitive scientists appeal to notions of representation in their explanations of cognition" [Ramsey, 2007, pxv]. As Tyler Burge puts it, there may be elements of a theory of representation which cognitive science, "without being fully aware of its own accomplishment" [Burge, 2010, p9], has discovered, and which the philosopher seeks to uncover.

Nico Orlandi provides the following characterisation:

> I propose that we look at what mental representations are by looking at how they have been used in these disciplines [i.e. cognitive science, broadly construed]. In this respect, I take philosophers interested in the notion of mental representation to be akin to those philosophers and historians of science more generally who investigate the nature of scientific posits by looking at scientific practice. [Orlandi, 2020, p101]

This methodology rests on the assumption that there *is* an implicit theory of representational content operative within cognitive science. Frances Egan argues that the assumption is false; cognitive science rests on no implicit theory of representational content. According to Egan, attributions of content in cognitive science are a "gloss", not part of the "theory proper". Content is attributed to a mental representation on the basis of pragmatic choices by the scientist; she chooses, from the many environmental items

which have some degree of co-variance with the representational state, the item which 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 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 which co-varies with the neural representation[1]. The only reason she chose ⟨*face*⟩, so Egan argues, is that it specifies the most intuitive environmental item, helping the reader understand the 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.

In the next section, 3.2.1, 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 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 help guide us towards those parts of cognitive science in which content attribution is guided by an implicit theory. If a scientific paper fails to meet these principles, we are likely to find that content is a gloss.

[1] An image of two dots above a curved line is a stimuli sometimes used in facial recognition tasks (e.g. [Tsao and Livingstone, 2008, p10]).

3.2.1 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. We can derive a non-deflationary theory of content from cognitive science only if content plays a role in the theory proper.

2. Content is part of the theory proper 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, p251].

   (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, p251].

3. Content is not part of the theory proper because:

   (c) Cognitive states and structures are not individuated in terms of content, rather they are individuated by the mathematical functions they compute.

   (d) "since pragmatic considerations typically *do* play a role in determining cognitive contents, these contents are not determined by a naturalistic relation" [Egan, 2018, p255]. 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. Since content does not play a role in the theory proper, we cannot derive a non-deflationary theory of content from cognitive science.

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, p253]. So, cognitive content must be part of an explanation of the functioning of the system in which the representational state which relates to that content is embedded.

Egan characterises cognitive content as the kind of content that "theories of content developed by philosophers" [Egan, 2018, 250] are aimed at. Certainly, for our purposes, this is the kind of content aimed at: the methodology under investigation is employed by those philosophers who are attempting to provide a theory of content consistent with the practices of cognitive science. Each adherent to the "explanatory turn" [Schulte, 2023, p55] is committed to cognitive content in Egan's sense. So, for the theorists we are targeting, content must play some explanatory role for the system in which its related representation is embedded. Going forward, all references to content specify cognitive content.

A **non-deflationary** account of content is an account in which content plays an explanatory role. As the argument suggests, 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, p119].

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:

> a specification of the function (in the mathematical sense) computed by the mechanism, specification of the algorithms, structures, and processes involved in the computation, as well as what I call the ecological component of the theory – typically, facts about robust covariations between tokenings of internal states and distal property instantiations under normal environmental conditions. [Egan, 2020, p11]

Egan maintains that these components together 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, p11]. We can therefore minimally characterise the theory proper as that element of the scientific theory which is sufficient to explain the cognitive capacity under investigation. This definition is neutral on the question of how to characterise a scientific explanation.

The theory proper is to be contrasted with the **gloss** cognitive scientists use when presenting their theory [Egan, 2018, p254]. 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, p12].

When looking to recover a naturalistic theory of content from cognitive science, the philosopher should hope that content is used within the theory proper. Otherwise, we will not be tapping into some theoretically informed usage, the principles of which we hope to uncover. Rather, if the cognitive scientist chooses from a plethora of possible content attributions based on optimising various heuristic values, we have a *pragmatic*, not *naturalistic* theory of content.

My response to Egan's argument will be to demonstrate that, in the case study I introduce, cognitive content is also required for a complete explanation of the cognitive capacity under investigation. Since this is the case, it can also be shown how the use of content in the case study meets the necessary conditions for inclusion in the theory proper set out in (a) and (b) in the above argument. Condition (a) is clarified in section 3.3, and (b) in section 3.4.

If we are to respond to Egan, we need to refute both (c) and (d). In what follows, 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 which 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.

3.3 Is content essential?

To be essential, cognitive scientists must "individuate the states and structures they posit[...] partly in terms of their content" [Egan, 2018, p251]. This entails that the content cannot change while the representation remains of the same type. Egan argues that mathematical content is essential, whereas cognitive content is not. The mathematical function might remain constant, while the cognitive content changes:

> If the mechanism characterized in mathematical terms by the theory were embedded differently in the organism, perhaps allowing it to sub-serve a different cognitive capacity, then the posited structures would be assigned different cognitive contents.[Egan, 2018, p255]

In contrast, the mathematical function computed remains the same regardless of external (or internal) changes. Discussing Marr's work on early visual processing, Egan writes:

> The mechanism described by Marr would compute the Laplacean of a Gaussian even if it were to appear (*per mirabile*) in an environment where light behaves very differently than it does on earth, or as part of an envatted brain. It would compute this function whether it is part of a visual system or an auditory system, in other words, independently of the environment—even the internal environment — in which it is normally embedded. [Egan, 2014, p122]

While the content of a state might change depending on external factors, the mathematical function it computes does not.

This calls for some clarification. It is not the case that the mathematical function computed by a cortical area remains the same no matter what. Neural plasticity describes the change 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 processes the Laplacean of a Gaussian) and place it in an entirely different system, a different mathematical function *may* in fact be computed. For example, if the input distribution is no longer Gaussian, the cortical cells' response profiles can be altered to efficiently encode the new statistical distribution of their inputs (e.g. [Laughlin, 1981], [Friston et al., 2006]).

However, Egan's point is, more precisely, that *were* the mathematical function computed to change the state itself would change, since the state is type-individuated with respect to the mathematical function it computes. Computing a mathematical function is *constitutive* of that state being that state. This suggests a flat-footed response to Egan: if the representation's content changes, we simply stipulate that the state has changed. We may follow Burge in maintaining that "the natures of many mental states are *constitutively dependent* on relations to the environment" ([Burge, 2010, p63]). Under this construal, the type of state something is depends on its relations to the external environment.

However, crucially, Egan argues that states can be constitutively dependent on x only if a change in x results in a change **for the system itself**. What constitutes a change for the system itself, and why does Egan argue that this is required for constitutive dependence?

Unfortunately, Egan provides no general criteria for assessing whether a change constitutes a change for the system itself. It is also unclear how we should understand the

system for which there is a change. There are a number of possibilities. We may define the system as the whole system of which the subsystem is a part. We may define the system as the subsystem itself. We can also think of subsystems themselves at various levels of grain. We might take the cortex, or the visual system, or V1, or specific pathways within V1, or the immediate surrounding structure of the representational state or structure under investigation.

It is also unclear whether a change in a fine-grained subsystem must lead to a change in the system as a whole. If there is some degree of redundancy in determining downstream processing or eventual behaviour, some local change may fail to produce a change at a higher level.

To remove ambiguity, I propose that we adopt the following characterisation of what it means for there to be a change for the system itself. In order for a change to be a change for the system itself, the change must affect either (a) **what** the the system does with respect to a given cognitive capacity (for example, face recognition) or (b) **how** the system performs that cognitive capacity. 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, 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 for the contribution to the cognitive capacity of the target representational state or structure (including a hypothesis at the mathematical level of description) must be invalidated or refined.

These two criteria involve changes for 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.

With (a) clarified, in the next section I attempt to counter (c) in Egan's argument. 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 in the system itself (in the sense specified above).

3.3.1 Case study: Chang and Tsao (2017)

In this section I present a case in which two representational states share a mathematical function, but are individuated with respect to their content. I claim that the researchers treat content as essential: they type individuate the two representational states partly in virtue of their differing content. Crucially, a change in this content results in a change in the system itself (argued for in detail in section 3.3.3).

Chang and Tsao [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.

Chang and Tsao discover two types on 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[2].

They suggest that each type of cell contributes 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 which require just one of the two types of content.

Chang and Tsao observe that "the fundamental difference between ML/MF [middle lateral/middle fundus of inferotemporal cortex (IT)] and AM [anterior medial of IT] lies *in the axes being encoded (shape versus shape-free appearance), not in the coding scheme*" [Chang and Tsao, 2017, p1020; 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" [Chang and Tsao, 2017, p1020] with the "incoming face" expressed as a 50-d vector and the "face space defined by the cell's STA" another vector which, roughly, characterises the response profile (STA) of the cell (for a detailed description of STA see section 7.4.2).

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 "spike-triggered average", of the cell is the *average stimulus* that the cell responds to [Chang and Tsao, 2017, p1015]. Before the STA is found, the stimulus range is "parameterized" [Chang and Tsao, 2017, p1022] along the dimensions of shape and appearance, divided into 25 shape metrics and 25 appearance metrics. Each new stimulus is generated

[2]We will investigate precisely how these terms are defined in section 3.4.2.

by software which randomly assigns values for each of these 50 parameters. We can 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 type-identified 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 3.4.2).

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 which 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* which determines the axis vector values is the observed response of the cell to the the corresponding content which 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, which 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.

Throughout the following sections I argue that we should consider this study to be a case in which cognitive content features in the theory proper.

3.3.2 Ecological component versus encoding

Egan allows what she calls the "ecological component" of the theory [Egan, 2018, p253] into the theory proper. The ecological component is an external item which the states or structures under investigation "typically correspond to" [Egan, 2018, p253]. The ecological component is given by "facts about robust covariations between tokenings of internal states and distal property instantiations under normal environmental conditions" where these covariations "constrain, but do not fully determine, the attribution of cognitive content" [Egan, 2020, 33].

The ecological component of the theory, given that it is specified by covariation, does not succeed in picking out just that aspect of the environment a change in which makes a difference for the system itself. Some aspect of the environment which covaries with a state may change, while another aspect of the environment, a change in which *would* affect a change in the system itself, remains the same. For example, a certain complex shape profile covaries with faces. However, some systems, such as AM cells in IT, respond to those complex shapes if they appear on toast, clouds in the sky, or a crude drawing. For the purposes of isolating just those properties which make a difference to the system itself, some relation other than covariance must be specified within the theory.

In what follows, I argue that Chang and Tsao do not rely on covariance relations to establish the content of a representation. Rather, they invoke, as is familiar in cognitive neuroscience, the relation of **encoding**. Multiple references to encoding are used throughout the study. The study itself is an investigation into what particular cells encode. Evidence that Chang and Tsao use encoding in the technical, information-theoretic, sense (i.e. not as a gloss on covariance) is given by their discussion of the particular encoding function used by the system, discussed below.

I first define encoding, then spell out two implications of the encoding relation which enable specification of content beyond covariance. First, the encoding relation requires a *function*[3] on the side of the system, relating upstream areas to downstream capacities (section 3.3.2). Second, the encoding relation requires that the system be able to *decode* the encoded input, placing constraints on the representation's surrounding architecture

[3]Not to be confused with mathematical function. When used without qualification, I use 'function' to specify the cognitive function of a representation - very broadly: the role performed, within a wider system, by a subsystem which enables the cognitive capacity which the subsystem serves.

(section 3.3.2). In section 3.3.3 I specify how, in virtue of these two features of encoding, a change in encoded content affects a change for the system itself.

Encoding and functional relations

Encoding, in essence, is the process of converting symbols. Here is a textbook definition of an encoder:

> Before being transmitted, each message **s** is transformed by an encoder, which we can represent as a generic function *g*, into the channel input $x = g(s)$, which is a sequence of *codewords*. [Stone, 2015, pp26-7]

A message is modelled as a value of a random variable[4]. We can generalise the definition of an encoder to produce the following definition of **encoding**:

> X encodes Y only iff there is some mathematical function f which 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 $x_{(1-n)}$.

In order for some system to encode an input, we must specify the values of the input to be encoded. In information theory as originally conceived by Claude Shannon, what is encoded is clear. Encoding was a notion intended for use in telecommunications [Shannon, 1948]. In telecommunications, we know what is being encoded given that we have designed the system ourselves. In telephones, strings of sounds from a person must

[4] A set of values with a probability distribution over them, denoted with a capital letter, e.g. X - see section 5.3.

be encoded as electrical signals, before being decoded as strings of sounds at the receiver. However, in natural systems we must construct theories about what is encoded. So, how do we specify precisely which values, or 'messages', have been encoded? I maintain that we must defer to the function of the system within which we find the encoder.

The role of an encoder as described by communication theory is to allow a sender-receiver mechanism to reproduce "at one point either exactly or approximately a message *selected* at another point" [Shannon, 1948, p379; emphasis added]. Determining which message has been 'selected' requires knowing the function of the encoder within the wider sender-receiver system. In simple terms, we need to think of the *use* of the encoded content[5]. It depends on the message which is meant to be reproduced at the receiver, which constrains which content is selected for encoding. In our terms, this means that what the downstream cognitive systems, which receive input from AM cells, *do* with the input from the content they receive matters for the characterisation of the encoding performed by the AM cells. It is not the scientist who selects a content from a range of covarying states; the *system itself* selects the content based on what it *needs* in order to perform its cognitive task.

Chang and Tsao hypothesise that $\langle shape \rangle$ and $\langle appearance \rangle$ are encoded, since "one can linearly decode [these] features and use these decoded features flexibly for any purpose, not only for face identification". They have in mind, specifically "tasks such as gender discrimination or recognition of daily changes in a familiar face" [Chang and Tsao, 2017, p1024]. When they say "one" can decode these features, they mean downstream systems;

[5]Speaking of encoded content avoids the circumlocution of referring to the messages which can be traced back to the source item. What is encoded, strictly speaking, is a set of signals which the content (source item) outputs (e.g. the light signals which bounce off the content proper, the shape).

"downstream areas [read] out the activity of AM with greater flexibility to discriminate along a variety of different dimensions" [Chang and Tsao, 2017, p1024].

The content of Am cells is characterised as $\langle shape \rangle$, understood in the technical sense to be outlined in 3.4.2, since this is the content which is suited to be used, generally, by the myriad downstream systems which perform a number of tasks.

Encoding and decoding

Chang and Tsao aim to uncover the *specific encoding strategy* used by the cognitive system under consideration. They develop the following theory: AM cells use a linear coding strategy to encode shape information. That is, AM cells respond in a linear fashion to shapes, with a higher firing rate as the shape deviates from the average shape stimulus which triggers the cell in one direction, and a lower firing rate as it deviates from the average in the opposite direction (e.g. when two points on a shape move further apart or closer together). The linear encoding strategy is posited in virtue of the fact that it provides a "simple" strategy for, as Chang and Tsao put it, projecting shape information onto the axis of the cell as defined by its STA [Chang and Tsao, 2017, p1022]. This is implemented, as we saw in section 3.3.1, by the cell (acting in such a way as can be modelled as) taking the dot product of two vectors, a relatively straightforward calculation.

Linear decoding is not a given. A non-linear encoding strategy can be more efficient, in the information-theoretic sense of preserving the most information possible between sender and receiver. If the distribution of the relevant features in the environment is Gaussian (as is the case in Chang and Tsao's study), the optimally efficient encoding of this information is achieved by using a cumulative density function (CDF). A CDF results

in an S-shape response profile; the cell has a greater sensitivity to differences among the most common inputs, and a lesser degree of sensitivity to inputs at the extremes of the Gaussian distribution. The cell 'cares' more about the most common inputs, distinguishing between them carefully, while treating less common inputs as more or less alike.

A non-linear decoding strategy is computationally demanding. A linear decoding strategy, while computationally simple, is lossy. Which decoding strategy is employed by downstream areas reflects a trade-off between information loss and processing simplicity.

When positing the content of a representation, we must consider what can be *decoded* from that representation. The decoding strategy is a an important aspect of this consideration, since it constrains the amount of information about the input available to downstream areas. This is a point emphasised by de-Wit et al. (2016) [de Wit et al., 2016]. They write:

> Much 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, p1415]

de-Wit et al.'s concern closely resembles Egan's; if de-Wit et al. 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, recovering more information about the input than the system itself, which can only perform linear decoding.

If we are interested in content as essential to the representational state under investigation, we must be concerned with what the system itself can decode. Otherwise, we attribute to the system representational content which it is incapable of using for the performance of downstream cognitive capacities. 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.

Chang and Tsao are explicit in their hypothesis that the system itself linearly decodes the input. Their study primarily focuses on what they, as researchers, could decode from the cells, but with the aim of demonstrating the feasibility of such a decoding strategy for the system itself - they "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, p1024]. 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, p1019].

The hypothesis, that the system itself is able to retrieve the same information that the researchers themselves were able to retrieve, requires further testing. Work is underway elsewhere in cognitive neuroscience, in the work of Philippe Schyns and colleagues (e.g. [Zhan et al., 2019]), to employ information-theoretic measures to perform such tests (see section 4.4.3).

For our purposes, the relevant point is that the input which is encoded should be considered hand-in-hand with what the system itself is hypothesised to be able to decode. We need to think of encoding and decoding as a coupled system, such that if some system encodes some input, that encoding is of input which is *for* the system itself, in the sense

of being retrievable. This is the case if we want to consider content as essential, and is precisely what Chang and Tsao, along with other cognitive neuroscientists, implicitly or explicitly practice.

3.3.3 A difference to the system itself

We have been considering the question whether content is essential to the representational states posited by Chang and Tsao. Essential content has two components: first, the states in question must be individuated with respect to that content, which is the case (section 3.3.1); second, individuation must be such that if the content with respect to which the states in question are individuated were to change, a change would result for the system itself. In section 3.3.2 we argued that the ecological component of the theory is not sufficient to isolate a content which meets this criterion. However, in section 3.3.2, we argued that Chang and Tsao do not invoke the ecological component, since they identify a relation other than covariance - encoding. In this section, I spell out how the two features of encoding detailed above ensure that a change in content results in a change to the system itself.

Let us remind ourselves of Egan's Marrian example:

> The mechanism described by Marr would compute the Laplacean of a Gaussian even if it were to appear (per mirabile) in an environment where light behaves very differently than it does on earth, or as part of an envatted brain. It would compute this function whether it is part of a visual system or an auditory system, in other words, independently of the environment —even the internal environment — in which it is normally embedded. [Egan, 2014,

p122]

Marr describes the content of the representational state which computes the Lalacean of a Gaussian as $\langle edge \rangle$. Generally, the system in question allows us to detect the edges of objects. Luminance changes typically occur at the edges of objects (i.e. between the object and its background), and the Laplacean function 'sharpens' the relatively gradual (Gaussian) change in luminance.

Under certain conditions, the content, $\langle edge \rangle$, changes while the mathematical function computed, i.e. taking the Laplacian of a Guassian distribution (in this case, of luminance levels), remains the same. For example, imagine that light is now reflected by objects in a very peculiar way: sharp changes in luminance are now found across physically continuous objects. Imagine a square table, all made of the same material. On one half, light behaves as it does in our universe and reflects as usual. On the other half, light, suddenly sensitive to the behaviour of the light to its left, decreases the intensity with which it is reflected. The overall effect is that luminance levels drop suddenly halfway across the table.

Similarly, at the edges of objects, there is now an effect whereby luminance levels remain constant. Again, now sensitive to its surrounding conditions, the light now increases intensity to match nearby light levels. We now have the following situation: sharp changes in luminance no longer correspond to edges of objects, but to their centres. Nothing has changed for the visual system: the same area computes the Laplacian of a Gaussian as though nothing has happened (or so Egan maintains).

There are a few possible situations we can imagine for the study we are considering. In all of them, a change in content either requires or affects a change in the system itself.

Below, I consider these imaginary situations. I describe how the two elements of encoding identified above ensure a change in the system itself in those situations, either a change in functions or a change in decoding.

Content change and function

Not *any* change in the environment relevant to the target content requires a change in content, even if content is explanatory. It is possible that the system comes to *mis*represent some content as present at some location where it is not. This will be the case in which some representation still has the function to present some information to a downstream area, but the environmental item which activates the representation has changed in a way which is inconsistent with the performance of that function (see section 6.6.2). Imagine a situation in which a sound starts beaming intermittently from space, targeted at specific individuals, uncorrelated with the presence of faces. When this sound is beamed at a person, shape-sensitive AM cells are triggered in the hearer. Imagine also that, in a completely unrelated series of events, all faces lose their shapes, leading everyone on the planet to look like a smooth mannequin.

What has become of the content of the target AM cell representations? In the short term, downstream face recognition systems still use shape input, and they will continue to respond to whatever input they get *as though* it picks out shapes. In this case, the content remains $\langle shape \rangle$. The target representations will now systematically misrepresent the space-sounds as shapes. There has been no content change at all.

So, how might the content change? There are two possibilities, one internal and one external. First, the function of the system containing the representation changes. We can

imagine that in our scenario, AM cells undergo a plastic neural wiring update and begin transmitting information to downstream areas which process sound for the purposes of some further cognitive task involving space-sound identification. Now, the content has changed, but this *requires* a change in the system itself, both in the cognitive capacity served, and in the local connectivity of the system containing the representation.

Alternatively, the content can change given a change in the environment, but one which is consistent with the performance of the function the system serves. Consider, instead of the space-sound being unrelated to faces, space-sounds now occur every time there is a face present. Specifically, we have a one-to-one mapping of space-sounds to the shape properties which previously existed on each face. Now, space-sounds *can* be used to recognise faces. In this scenario, we may have a case in which the content has changed, since the source item is now $\langle$*space-sound*$\rangle$. The function of the system - to enable myriad tasks, which we previously (section 3.3.1) saw required the positing of shape content - is now consistent with space-sound content. This is a case in which the new content falls within the functional profile of the subsystem containing the representation (see section 6.5).

Is this a change in content in which there is no change for the system itself? In the next section I argue that it either is not, or if it is, is precisely the same situation in which we might change the mathematical function computed without a change for the system itself.

Content change and decoding

If space-sounds enable face recognition, we can provide a theory for precisely how that happens. We can investigate which decoding strategy is used, and how the space-sound properties can serve downstream areas just as well as the previous shape properties. It is no small thing that space-sounds can (a) trigger responsivity in AM cells and (b) do so in precisely the way in which shape properties used to, enabling facial recognition. Specifically, we need to know precisely how nearby downstream systems can decode the very information they need to perform the myriad recognition tasks AM cells previously engaged in. We also need to know precisely what the encoding scheme is for this new information. Both encoding and decoding are likely to drastically change, leading to a difference in the system itself at various levels of analysis - including the level of the ascription of mathematical function computed.

Alternatively, the surrounding systems remain the same, and no new encoding or decoding strategy is required. However, this is only possible given that (aside from the fact sound would have to, magically, perfectly manipulate the AM cells directly, circumventing auditory processing channels) the space-sound profile perfectly overlaps with the shape profile, such that the former is isomorphic to the latter. If we can apply exactly the same encoding scheme, with the same values in each of the same vector positions, and the same linear decoding scheme with no loss of information relative to shape processing, with the same level of performance of the cognitive capacity under investigation, the change in content affects no change in the system itself.

However, in the case of perfect isomorphism, a change in mathematical function computed *also* makes no difference to the system itself. If we take a mathematical function

isomorphic to taking the dot product of two vectors, of which there are potentially an infinite number, we can ascribe any one of those mathematical functions, differences between which make no difference to the system itself. So, we might amend our original definition from section 3.3: a representational state is constitutively dependent on x only if a change in x results in a change for the system itself *up to perfect isomorphism*. This reflects an argument we will pick up in the next section: for some range of external items, natural indeterminacy is to be expected. If a change in content makes absolutely no difference to internal processing, or the cognitive capacity under investigation, the representation may be truly *indeterminate* with respect to a range of contents, or with respect to a range of mathematical functions.

In summary, I have argued that encoding and decoding ensure that a change in content reflects a change in the system itself, since the decoding scheme must change to accommodate new inputs, even if the same capacity is realised in the same way. If the same capacity is not realised in the same way, this is a clear case of content leading to a change in the system itself (given the definition in section 3.3). If the same capacity is realised in the same way, *and* no new decoding scheme needs to be enacted, the content is isomorphic with the previous content - but then change in content makes no difference for the system itself in precisely the same way that change in mathematical function computed results in no difference for the system itself. This is a case of natural indeterminacy, and perfectly scientifically acceptable. In the next section I will explore how the system reduces indeterminacy to this acceptable level, and how scientists reflect that in their theories by using technical vocabulary to isolate content.

3.4 Is content sufficiently determinate?

This section offers an argument against premise (d) 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, p255]. 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.

While we lack consensus on a general notion of what it takes for something to be naturalistic[6] the relevant comparison for our purposes is pragmatism in Egan's sense. So, to be part of the theory proper content must be determined by principles which do not invoke communicative heuristic values. Indeed, the relation should be spelled out in terms which 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 3.4.1, decoding provides a scientifically acceptable way in which a sufficiently determinate content is selected by the system itself.

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 explanation. 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 theoris-

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

ing except insofar as they violate this requirement. Content need be only as determinate as any other scientific posit.

I will not enter into a debate about whether content must be determinate in an absolute sense in order to be properly considered content. Perhaps indeterminacy is a natural property of representational content. Indeed, Karl Bergman argues for precisely this point [Bergman, 2023]. Minimally, we should allow for the *possibility* that some representations fail to specify just *one* item under just *one* description.

Rather, we should 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. What is required to achieve this level of determinacy? I first argue that, on the side of the system, the decoding constraint provides a system-side limit to the set of possible distal contents. I then argue that, for the purposes of informational teleosemantic methodology, we must find studies which use technical terminology to isolate content, rather than rely on intuitively grasped ordinary terms.

3.4.1 Constraints due to coding

For any given function, there are a number of environmental items which could, in principle, fulfill that function. We saw an example in section 3.3.3. So, how does the *system itself* determine which item *will* fulfill that function? In short, the answer is coding constraints, on both the input and output side. Chang and Tsao's use of the encoding relation implicitly imports these constraints on content. In doing so, they are able to rule out contents which have no explanatory value.

I hope to show that the constraints implicitly imposed by the encoding relation allow

us to rule out a swathe of indeterminacy, narrowing in on contents which can be detected by the system itself, as well as utilised by downstream systems. This points towards sufficiently determinate content, given by naturalistic system-side constraints.

What the system can *encode* places constraints on representational content. For example, if some channel can only process visual information, only visual information can be encoded. As we saw in section 2.4.1, Neander argues 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 $\langle green \rangle$ content is not actually $\langle grue \rangle$? Grue is defined as "(i) seen before 2040 and green or (ii) seen after 2040 and blue" [**?**, p169]. 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, 2017b, p169]

Constraints on what the system can discriminate in the environment help us reduce the range of possible environmental items which serve as content for a given representation. Representations with either $\langle$*green*$\rangle$ or $\langle$*grue*$\rangle$ content can fulfill a number of functions which *prima facie* call for $\langle$*green*$\rangle$ content (until 2040, at least). Nonetheless, if the system has no way of detecting grue, grue cannot be the content of the representation. I consider this point in more detail in section 5.3.1.

In addition to upstream encoding constraints, what the system can *decode* places constraints on representational content. What the system itself can use from the representa-

tion limits the range of possible environmental items the representation represents. This was discussed in section 3.3.2 with reference to de-Wit et al. [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 which 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 the 1st of January, 2040.

However, imagine the system itself has downstream neurons which fire in response to the input of the colour-sensitive neuron. Imagine also that not a single one 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 this provides a plausible naturalistic principle for limiting the range of possible content ascriptions. Moreover, the constraint appears to be implicit in the study under investigation. As I argued in section 3.3.2, 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 in the system itself, and we have good reason to think that such a limitation provides us with a range of content ascriptions which are explanatorily relevant. Coding limitations, encoding and decoding, are the system's own way of reducing indeterminacy to within acceptable levels.

3.4.2 The use of 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. 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.

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

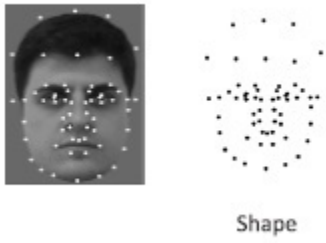
It should be clear that using the *ordinary* (non-technical) concept SHAPE does not serve Chang and Tsao's explanatory purpose very well. There are numerous types of shape, and the concept is very broad. The system under investigation, as I have emphasised throughout, has as its content a very specific 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 investi-

gation.

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. Below, I begin by setting out the process by which Chang and Tsao design their target phenomena, which they identify with the term 'shape'. It is shape understood in this technical sense which determines the precise content of the AM cells.

In order to arrive at their target phenomena, first "a set of landmarks were labeled by hand" on images of 200 faces from "an online face database" [Chang and Tsao, 2017, p1015], which can be seen in figure 3.1.

Figure 3.1: Labelling landmarks by hand.



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" [Chang and Tsao, 2017, p1015]. Principal components analysis (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 which 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" [Chang and Tsao, 2017, p1015].

In terms specific to shape, 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 which define the bottom (which we could call the 'chin', for example). I will attempt to justify this deflationary description below, but for now it should help us to clarify how the shape content can be minimally characterised.

Each individual cell is thought to encode, on average, "6.1 feature dimensions" [Chang and Tsao, 2017, p1015] 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.

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 which they identify. They provide a description of the specific encoding scheme enacted by the AM cell, relative to these parameters, which explains the behaviour of the cell.

As may be apparent, throughout I have resisted the temptation to assert that any of the representations in question have contents such as $\langle face \rangle$. I have persistently asserted

that the ⟨*shape*⟩ 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, 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 which 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 which theorists such as Chang and Tsao are investigating.

What is the moral of this section? The distal item which is the content of the representations under investigation has been specified using technical terminology, isolating relevant phenomena to a far greater degree of specificity than non-technical language allows. Further, the phenomena isolated are explanatorily relevant to the cognitive capacity under investigation. As such, the indeterminacy which 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 which does not isolate the same relevant phenomena, without loss of explanatory power.

It will help to bring out the distinction between technical and non-technical language

by considering how EDGE as used in cognitive science differs from how the term 'edge' is used in non-technical language. It is debateable, when considering intuitions regarding non-technical language, whether perfectly round cylinders have edges along their curved sides. Can something which is smoothly round have an edge? However, the technical concept EDGE as used by Marr is not susceptible to such debates. The limit of the cylinder from one's perspective constitutes an edge, within the technical language of the theory. As used in this theory, the technical concept EDGE isolates limits of objects, whether or not those limits correspond to a sharp edge. The technical use also eliminates the polysemous nature of the non-technical term - one can be 'edgy' or have an 'edge' in the non-technical use of the term, but it is obvious that the technical concept does not isolate whatever phenomena make one 'edgy'.

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, we have the same representational content. In other words, such purported indeterminacy is a difference which makes no difference. By employing their technical language, Chang and Tsao ensure that the content they attribute to representations is sufficiently determinate so as to be irreplaceable within the theory, except by language which isolates just the same phenomena. At this point, any difference is merely terminological, with no bearing on the distal item which constitutes the content proper.

3.5 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 which use content within the theory proper.

3.5.1 First principle

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

As Shea writes, if we have an externalist explanandum, we will have an externalist explanans [Shea, 2018, p31]. As we saw in section 3.3.3, if some representation has a function to serve a cognitive capacity which requires interaction with the external environment, the external item which is required for that capacity to be realised will be the content of the representation even in cases of misrepresentation. A change in content *requires* that the function changes, provided the new content does not also suffice to enable the successful functioning which the original content enabled.

A function can be either explicit or implicit, entailed by the positing of the encoding relation. However, encoding must be used in a transparently technical sense in order for us to be sure that there is an implicit theory behind *its* use.

3.5.2 Second principle

Focus on studies which provide a hypothesis about what the system itself can decode, or otherwise access.

As we saw in sections 3.3.3 and 3.4.1, positing a possible decoding of content is crucial. It is required for a function-sustaining 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.

3.5.3 Third principle

Focus on studies which 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 which 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.

3.6 Conclusion

We must be vigilant when approaching cognitive science looking for a theory of content. Our methodology is secure, but we must be careful in those studies we apply it to. Egan is right that we have to treat content as essential, and provided by a sufficiently determinate naturalistic principle.

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 relation of encoding. Encoding ensures that there is a function served by the content, and that downstream systems can decode that content. A change in the distal item, triggering the representation, which does not serve the function, does not lead to a change in content. Instead, we get systematic misrepresentation, likely with disastrous effects for the whole organism. A change in the distal item which is consistent with performing the same function can lead to a change in content. However, this change in content either affects a change in the system itself - by way of altering the encoding and decoding strategy of the system - or it is perfectly isomorphic and is equivalent to changing between isomorphic mathematical functions, which also results in no change for the system itself. For good reason - some degree of natural indeterminacy is perfectly possible and should not be ruled out *a priori*.

When it comes to reducing indeterminacy to acceptable levels, the system itself imposes constraints on representational content by way of coding constraints. In our theory, we must use technical terminology to isolate content in order to ensure that we remove indeterminacy resulting from natural language, and isolate just those properties which are explanatory of the cognitive capacity under investigation.

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

In the next chapter, I look at some contemporary studies which meet the three principles outlined above. They seek to explicitly address the kind of concern Egan raises with respect to content being for the system itself. I argue that, in those studies, information theory provides the background theoretical framework. I provide some examples of how information theory can be used to isolate contents which feature in the theory proper.

Chapter 4

Background Theoretical Framework

4.1 Introduction

In this chapter, I argue that information theory provides the background theoretical framework for content attribution in those regions of cognitive science which meet the requirements identified in the previous chapter. Rather than cite mere correlations, information theory provides a model of the precise interactions between the cognitive system and the environment which deliver essential and naturalistic content attributions.

I begin in section 4.3 by outlining what a background theoretical framework is in general. I describe some common background theoretical frameworks from multiple domains within cognitive science.

Then, in section 4.4, I show how information theory is used in the types of studies identified in the previous chapter. I argue that, in those areas of cognitive science which explicitly address the kind of concerns raised by Egan, information theory provides the

background theoretical framework for identifying contents. In essence, information theory allows one to trace information flows between the environment and the system (section 4.4.3), as well as within the system itself (section 4.4.4). Information theory provides a precise, mathematical model of the specific element of the environment which makes a difference to the system itself.

The chapter also serves as an explication of the power and complexity of information theory as applied to the cognitive system, in contrast to what I considered to be the presumed simplicity of content attributions in cognitive science in chapter two (section 2.7). I also prepare the way for later chapters by showing how the relation of maximal mutual information is taken to be the relevant informational relation in determining content.

4.2 Information in cognitive science

One may consider it a truism that information theory provides the background theoretical framework of cognitive science, since the language of information theory is rife within the field. For instance, a foundational text such as Marr's *Vision* invokes communication channels [Marr, 2010, p10], information processing [Marr, 2010, 19], efficient encoding [Marr, 2010, p262], noise reduction [Marr, 2010, p71] and so on. In general, since the cognitive revolution, cognitive scientists have generally considered the mind in information-processing terms.

However, we saw in the previous chapter that an appeal to cognitive science must involve more than observing linguistic practices. We cannot guarantee that the language used is due to any theoretical commitments, either implicit or explicit. It might be a

gloss. We have to be sure that the concepts used are doing some genuine theoretical work. By showing information theory in action, I demonstrate that information theory can be genuinely explanatory for modelling cognitive system/world relations.

4.3 What is a background theoretical framework?

As I will use the the phrase, a **background theoretical framework** for any domain consists in a model of the properties and processes in that domain, which generalise across lower-level properties and processes which underlie that domain (e.g. physical interactions, chemical synthesis, etc.) using a set of concepts and principles which are independently well understood.

Generally, cognitive science seeks to explain some phenomena by means of an independently well-understood set of principles forming a conceptual scheme. For example, at a very general level one may seek to explain behaviour in terms of intentions and desires (e.g. [Davidson, 2001]). One may further explain the connection between intentions and action in plan-theoretic terms (e.g. [Bratman, 1993]). In this case, theories of planning form the background theoretical framework for the account.

In a different cognitive domain, one may wish to explain the function of the visual system in terms of high-dimensional object manifolds (e.g. [DiCarlo and Cox, 2007], [DiCarlo et al., 2012]). Mathematical tools associated with this formalism form the background theoretical framework for the explanation of the behaviour of the visual system.

Alternatively, one may wish to explain the structure of stored knowledge in terms of neural networks, employing the background theoretical framework of connectionism

(e.g. [Hinton, 1990]).

Studies on face recognition and related capacities (tracking changes in faces, interpreting emotions using faces, identifying general features such as gender) employ the background theoretical framework of the face-space model (e.g. [O'Toole, 2011]).

A background theoretical framework provides multi-purpose tools, often mathematical or geometrical, to specify some aspect of the system under investigation (e.g. operations in terms of computations performed) or otherwise make claims about the system in virtue of how it is constrained by the framework employed (e.g. for Bratman, if plan theoretic tools can be used to model intentions, intentions must conform to, and be able to implement, rational constraints on planning as understood within that framework).

I argue that information theory provides the background theoretical framework of content attribution (at least for those regions which meet Egan's constraints). This is initially motivated by our guiding hypothesis based on previous work in informational teleosemantics, as set out in chapter two. However, the real argument for the position is provided just by looking at scientific practice and noting how information theory is actually used.

In the next section I introduce information theory and provide examples in which it supplies the background theoretical framework for content determination in cognitive science.

4.4 Examples of information theory

4.4.1 Outline of information theory

Information theory, broadly, is a mathematical tool which models relations between entities. It provides a measure of the statistical relationship between two or more entities, modelled as 'random variables' (see section 5.3). Information theory describes how many symbols (such as the symbols 1 and 0) are required to reduce a set amount of probabilistic uncertainty about some random variable, where each symbol reduces the number of possible states by half. For example, if you flip a coin behind a barrier I am uncertain about the result. However, since there are only two possible outcomes, information theory tells me that I only need one symbol in order to know which outcome obtained. The amount of information which one can obtain from a source is given by its **entropy**:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log p(x_i) \tag{4.1}$$

Where the log is taken to base 2 in order to measure the entropy in *bits*. One bit of information, given the use of log_2, reduces uncertainty about the source by half.

Information measured between two random variables with entropies $H(X)$ and $H(Y)$ is called **mutual information**. An example definition, in information-theoretic formalism, is:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{4.2}$$

Intuitively, the mutual information between two random variables involves the un-

certainty we have about each variable once we have rid ourselves of the uncertainty surrounding both variables together. Alternatively, it is quantitatively equivalent to the remaining uncertainty we have about a variable once we know about another variable.

4.4.2 Information theory in practice

In this section I primarily focus on work within psychophysics and related neuroscientific literature. Psychophysics deals with the interaction between external stimuli and internal sensory processing. At the most fundamental, psychophysics covers transduction - the conversion of one signal into another. This concerns, for instance, the behaviour of photoreceptors in the presence of electromagnetic energy, or chains of neurons connected via synapses. However, psychophysics also concerns macro-level interactions between sensory systems and the wider cognitive system.

I have chosen to focus on this work precisely because it concerns the interface between the cognitive system and the external stimulus. As we saw in our response to Egan, it is crucial to find work which explicitly looks at this interface, and which is careful about precisifying exactly which components of the stimulus are thought to serve as content. Additionally, as we shall see, the psychophysics literature increasingly emphasises the role of downstream systems in information processing. This was another crucial component in our response to Egan. As such, this work meets the criteria I set out in the previous chapter for finding a role for content within the theory proper.

We will see that there are many statistical models which can be employed to describe internal (e.g. between systems) and external (e.g. between an environmental item and sensory system) relations. However, I will follow the authors I cite in arguing for

the utility of information theory as a framework. Information theory has proved to be highly valuable in a number of ways, including in precise specification of environmental variables, discovering close relationships between internal systems, and discovering the brain locations of information integration.

It is true that other statistical methods are equally as powerful. Arguments for the application of information theory are therefore often pragmatic in nature; for example, information theory provides equations which are easier to compute than equivalent statistical methods. However, we are looking for a naturalistic theory of content determination - one which is not dictated by pragmatic considerations. In the next chapter I will attempt to show how, despite information theory being selected (partly) for pragmatic reasons, it can in any case be given a non-pragmatic interpretation based on features of the cognitive system.

The aim of the present chapter is more modest: we aim to understand whether and (if so) how information theory is (or can be) deployed to model links between distal items and the cognitive system, and between elements of the cognitive system.

4.4.3 Interface between stimuli and sensory systems

Studies which focus on the relation between the distal environment and sensory systems typically ask: *which stimulus drives a neuronal response?* In this section we see how information theory is used to model this relationship to a high degree of specificity.

As Fodor emphasises [Fodor, 1987], naturalism minimally involves using theoretical constructs which do not presuppose representational content (on pain of circularity). So, when attempting to see how information theory can be used in a naturalistic model repre-

sentational content, we should look for relations which are not already representational in nature. Causal interactions involved in driving a neuronal response provide a basic relation from which we can build a theory of content.

I will further break this section into two related areas dealing with the link between the cognitive system and the distal environment. First, I will describe how information theory can be used to model the precise elements of a stimulus which are responsible for driving a neuronal response - using the measure of **conditional mutual information**. Given its usefulness in this context, information theory can be deployed to tackle the difficulties which arise when studying relations to natural stimuli (i.e. those stimuli not explicitly designed by experimenters using controlled parameters). Second, I will describe how information theory can be used to model sensory adaptation (i.e. changes in sensory systems as a response to changes in external states of affairs).

Conditional mutual information

As described by Ince et al., conditional mutual information (CMI) "quantifies the relationship between two variables while removing any effect of a third variable" [Ince et al., 2017, p1549]. It is typically expressed as $I(X;Y|Z)$ where Z is the variable one is conditioning out, or removing the effect of. Ince et al. describe the potential application of this measure as follows:

> With many types of naturalistic stimuli, extracted stimulus features are highly correlated (for example, luminance of neighboring pixels of a natural image or the acoustic features of speech). Given an analysis of each feature alone, it is difficult to determine whether a specific feature is genuinely encoded in

> a neural response, or whether the response is actually modulated by a different correlated stimulus feature. CMI provides a rigorous way to address this issue. (p1549)

We can think about conditional mutual information as the information still available about X from Y once we already know Z. If Z 'fully explains' X, we will receive no more information about X by looking at Y. We can see this from one definition of CMI[1]:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{4.3}$$

Where $H(X|Z)$ and $H(X|Y,Z)$ are the conditional entropy of X on Z and the conditional entropy of X on Y and Z, respectively. When we say that Z fully explains X we mean that $H(X|Z) = 0$, such that the entropy of X is completely reduced given Z. In other words, there is no more information left to be extracted from X once Z is known. When this is the case, it is also true that $H(X|Y,Z) = 0$, but in this case the addition of Y is redundant.

So, in the situation where $H(X|Z) = 0$, $I(X;Y|Z) = 0$. Of course, this is an extreme case in which Z entirely reduces the entropy of X. Other situations will appear in which either of the following hold:

$$I(X;Y|Z) < I(X;Y) \tag{4.4}$$

$$I(X;Y|Z) > I(X;Y) \tag{4.5}$$

[1] A very useful guide to this topic is provided in [Cover and Thomas, 1999]. Further information (and an example of an application of the measure) can be found in [Renner and Maurer, 2002].

Which is used to define **interaction information** (the mutual information between three variables) as:

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) \tag{4.6}$$

This allows us to describe some situations in which CMI can be very helpful in discovering the explanatory relationship between X, Y and Z. For instance, in a situation in which equation (4.4) holds, the interaction information in equation (4.6) will be **positive**. This is true except in the case in which $I(X;Y|Z) = I(X;Y)$ when $H(X|Z) = 0$ and so $I(X;Y;Z) = 0$, which, since we know $I(X;Z) > 0$, indicates that there is no further information shared between the three variables over and above what is shared by two of them. Otherwise, if we have an inequality as in (4.4) we know that Z *partly* accounts for X over and above the involvement of Y.

Imagine that I want to explain the fact that I spilled my coffee. I can cite the fact that I very rapidly moved my arm to the left while holding my cup filled to the brim with coffee. If you learn that fact, (and based on your intuitive understanding of physics) you gain information that I spilled my coffee. However, imagine that someone pushed my left arm forcefully. This partially accounts for why I spilled my coffee. Perhaps if someone forcefully pushed my arm I could hold firm and not spill my coffee, so the explanation is not full. Nonetheless, one has at least some information about my spilling coffee from this third fact.

If, however, equation (4.5) holds, the interaction information in (4.6) will be **negative**. This can be described as a situation in which adding a third variable actually increases the amount of information one has about X from an observation of Y.

Imagine you know that I will only phone you under two circumstances: either I find

your missing cat or (I have insider info) I know you have secured a promotion. Assume that your missing cat being found and your promotion are statistically independent, such that if you know your cat has been found, you still have no information at all about getting a promotion (and *vice versa*). The phone is ringing, and you see it is me calling. Before you answer, an email appears telling you that you have failed to get the promotion. You are now in possession of the information that your cat has been found! In this case, conditioning on my call has increased the information available from knowledge about getting a promotion.

Why is this important? When applied to elements of a stimulus which drive neuronal responses, we can begin to tease apart exactly which elements are responsible. The world is a mess of statistically related elements, and redundancy abounds. If we wish to pinpoint just those features which drive a neuronal response, we can see if conditioning out other variables produces a lower CMI than MI. If so, that conditioned response has some influence on driving the neuronal response. If, however, the CMI score is the same as the MI score, we know the variable we condition on is statistically 'screened off' from driving the neuronal response.

This is helpful when experimenting with natural stimuli. However, the measure also has some surprising applications which (informally) look closer to something like pinpointing content. I provide an example to illustrate how we can begin to use these measures more broadly. Bröhl et al. (2022) [Bröhl et al., 2022] set out to investigate information gained in lip reading. Specifically, they attempt to answer whether "temporal and occipital cortices represent auditory and visual speech features during lip reading" [Bröhl et al., 2022, p7].

In order to answer this, Bröhl et al. first define the features they are looking for representation of. They define three auditory features (AudFeat): the signal envelope of the sound of a spoken word (i.e. the range of frequencies which that signal occupies), the slope of the sound (whether it rises or falls), and the dominant pitch of the sound. They define three physical features of the lips (LipFeat): the area of the lip opening, the slope of the lip, and the width of the lip opening. The question they ask is: which feature is represented within the given regions of interest (ROIs)?

Participants are shown either a silent video of speech (which included the whole face), or an auditory recording of speech without visual input. We will narrow our focus on just the results from the video-only trials, since they are perhaps the most surprising and demonstrate the potential of CMI. Using a measure of mutual information with activation of the region of interest, Bröhl et al. found that mutual information between AudFeat and the temporal ROI was present. More importantly, using CMI, they found that "the temporal ROI tracks the unheard AudFeat to a similar degree as when discounting the actually presented visual signal" since "there were no significant differences between MI and CMI values" [Bröhl et al., 2022, p8].

As we saw above, in a case when $I(X;Y|Z) = I(X;Y)$ we can say that there is no additional information which is gained from Y about X which is not already included in Z. If we take activation in temporal ROI as X, LipFeat as Y and AudFeat as Z, we can say that the features one is representing are AudFeats, even when one only has access to silent video. That is, there is no more information gained about the neuronal response given the LipFeats versus the information one already has given the AudFeats. Here is what the experimenters conclude:

> the auditory and visual pathways are also capable of apparent 'restoring' information about an absent modality-specific speech component; while seeing a silent speaker, both auditory and visual cortices track the temporal dynamics of the speech envelope and the pitch contour respectively, in a manner that is independent of the physically visible lip movements. [Bröhl et al., 2022, p11]

In summary, when dealing with the interface between the distal environment and the cognitive system, information theoretic tools allow close specification of the element of the distal item that drives a neuronal response. This can have surprising consequences, as it can demonstrate that some non-perceived features of the environmental item can be decoded by internal mechanisms 'filling in the gaps' and providing the otherwise lost information.

Modelling a neuronal response profile is just one way to use information-theoretic tools. In the next section we will look at interfacing between internal systems. Here, we find work on tracing the flow of information through a system. Ultimately, through an exploration of these topics, we will flesh out the specifics of the information-theoretic framework we hypothesise to be present behind content attributions. We will also discover vital resources for providing a theory of content using information theory.

4.4.4 Interface between internal systems

In section 4.4.3 we investigated how information theory can be used to model the specific element of a stimulus which is responsible for driving a neuronal response. We saw that CMI is a measure which provides such a model. In this section, we will pursue the flow

of information further into the cognitive system. We can formulate the topic of this section as a question: once responsivity, modelled using information theoretic resources, has been established between some aspect of an environmental item and some neural activity, can we use information theory to model the downstream processes driven by that neuronal activity?

We would like to be able to do this. As mentioned in the previous chapter, we are interested in the constraints imposed on information processing by downstream systems. We have argued that the content of a representation must be determined by the information which can be extracted by the system itself, since this provides answers to how-questions. This requires knowing the internal connectivity of the system: we need to know which later systems use the information provided in order to know what they decode from it.

I will continue to rely on the work of Schyns and colleagues. To my knowledge, the work coming from Schyns' lab represents the most thorough application of information theoretic models to the cognitive system to date. The implicit theory of content determination I hypothesise to be behind content attributions in cognitive science more generally is most explicit in this work.

The paper I will focus on from Schyns' lab by Zhan et al. [Zhan et al., 2019], has the additional benefit of being an explicit response to worries which are very similar to those of Egan. Zhan et al. describe the worry [Zhan et al., 2019, p324], presented by de-Wit et al. [de Wit et al., 2016], that when employing information theory, neuroscientists do so without regard for whether the information attributed to some neuronal activity is a measure of what the cognitive system *itself* uses or what we as external observers are

able to measure. As de-Wit et al. write, experimenters discuss what is encoded "without ever providing evidence that those recorded responses reflect differences in activity that can actually be used (received or decoded) by other areas of the brain" [de Wit et al., 2016, p1415].

What experimenters can decode from neural firing can, at least in principle, differ from what that cognitive system itself can decode. When experimenters see a neuron firing in the presence of some stimulus, it is natural to assume the cognitive system itself receives information from that neuron about the stimulus. However, without tracing the internal causal or probabilistic interactions, it is not possible to conclusively determine whether this is the case.

We can re-cast this in terms of our response to Egan: we suggested that one key to ensuring that content plays a role in the theory proper is to discover internal constraints which (a) make a difference to the system itself, and (b) place naturalistic (i.e. non-pragmatic or interest-relative) limits on the range of possible co-varying entities which could serve as content. If we are able to trace internal connections and find which information is actually *used* by the system, we can provide support for (a). By limiting the range of possible items to those which are explanatory of output we can also build support for (b). As de-Wit et al. aptly put it: "for information to truly be information, it has to be a difference that makes a difference to a receiver" [de Wit et al., 2016, p1416].

The parallels between Egan's concern and the challenge from de-Wit et al. can be pushed further. They argue that early studies in edge-detection (e.g. [Hubel and Wiesel, 1959]), the mechanism targeted by Egan as an example (e.g. in [Egan, 2020], [Egan, 2018]), should be read as finding a mere *correlation* between "cells that would fire" and "stim-

uli that we would call edges" [de Wit et al., 2016, 1417]. This maps to Egan's concern that the underlying relationship is merely one of co-variance. As a result, de-Wit et al. issue a challenge to neuroscience: "we need to focus on the 'cortex-as-receiver' to track the causal dynamics from one area to the next to establish whether a measured response is indeed information used by the rest of the brain" [de Wit et al., 2016, p1418].

In order to respond to de-Wit et al.'s challenge, Zhan et al. set out to find the specific information which drives a behavioural response to some input, given some neuronal activity. Specifically, they trace the "causal dynamics from one area to the next" using information-theoretic measures. In this way, the experimenters hope to provide a hypothesis for the information which is used by the system itself.

It should be noted that it may be that there is some information which can be used to support the same behaviour which we have failed to account for. This would be to question the information the researchers hypothesise to be used by the system itself. The hypothesis about the information used to generate the behaviour might be wrong. This should not worry us too much. As an empirical research program, the specifics of the information processed by the system itself will always be open to these concerns. It is my hope that we will be in a better position to adjudicate such questions once we have an explicit theory of content determination on the table.

Redundancy and useful information

Zhan et al. performed a study with five participants. Participants were shown stimuli constructed from a painting by Dali, an ambiguous image in which one can either see two nuns or the face of Voltaire (see figure 4.1).

Figure 4.1: Dali's *Slave Market with Disappearing Bust of Voltaire.*

The painting was broken down into various 'information samples': areas of the original painting were isolated using a 'bubble' technique, and then separated into high or low spatial frequency ranges. Participants were then shown images composed of a number of bubbles of varying spatial frequencies.

Upon seeing each image, participants were asked to respond with either "nuns", "Voltaire" or "don't know". During the task, Zhan and colleagues recorded brain activity using MEG imaging. By taking an average over each image composed of various spatial frequencies and image segments, it was possible to correlate MEG activity with specific regions and spatial frequency bands for the original image. It was also possible to correlate MEG activity with each 'perceptual decision' (i.e. which of the three options the participant selected). In other words, specific parts of the image which elicited MEG activity was found, and the specific MEG activity which contributed to a decision was found.

Using these measures, Zhan et al. calculated the amount of Mutual Information (MI) between each "voxel" (i.e. a three-dimensional unit square of the brain) of MEG activity and the "information sample" (i.e. part of the image). They also calculated the MI between each voxel and decision (i.e. either "nuns", "Voltaire", or "don't know"). Each voxel was recorded over 400ms, every 2ms, in order to provide a measure of MI per voxel over time. The measure of MI for each voxel and sample was taken to show the "strength of feature representation" where the maximum MI value was taken to show the "maximum representation curve" over time [Zhan et al., 2019, e4]. [2] Similarly, maximum MI values were found between voxels and decisions over time.

Readings over time were taken to assess the variation in voxel activity as downstream areas became increasingly engaged. As activity propagated throughout the cortex (along both the ventral and occipital pathways) measures of MI were taken for each voxel, and it was found that MI increased successively over time for voxels along each pathway.

To anticipate the results, voxels which had information about the image (i.e. maximum MI with some sample or "feature") *and* which contributed to the decision (i.e. maximum MI with a decision) appeared to propagate along the ventral stream, whereas voxels with only information about the image (i.e. only maximum MI with some "feature") appear to propagate along the occipital stream and then "die out". As the experimenters put it: "a spatio-temporal junction exists between the occipital and occipito-ventral cortex around 170 ms, after which only behaviorally relevant features flow into the temporal cortex, with the processing of irrelevant features ending in the occipital cortex" (p321).

This analysis was built on information theoretic measures which are not quite as

[2] A full treatment of why maximum MI was taken to be the relevant measure - apparently defining the representational content (the "feature" of the image represented) - will be given in chapter seven.

straightforward as taking MI values (although MI is a component unit of measurement). Rather, a measure which the experimenters called "redundancy" was used. In intuitive terms, redundancy measures the amount of information in a voxel which is about image features, and which tells *us* (as observers) nothing over and above what we can know about the decision made given the presence of image features alone. If we were to analyse the features contained in an image, and measure decisions based on those images, we would be able to correlate the presence of image features with the decision. We could then predict which decision was made given some image features, and vice versa. Up to a point, we would get no more information from looking at MEG recordings about an image given that we know the decision, and no more information about the decision by looking at the MEG recordings given that we know the features. The thought is that "redundancy" specifies the information about the image which the decision making process *uses* to reach a decision.

Of course, redundancy is just one way of describing this measure, one which does not appear to help our ultimate case for realism; it looks as though the measurement is relative to what *we* as experimenters can observe. Nonetheless, there are interpretations which demonstrate the way in which what is being measured is intrinsic to the object of study. As the experimenters describe it, the measure provides the "set-theoretic relationship between three entropies" [Zhan et al., 2019, pe4]. However, the point will be much clearer if we use familiar measures to describe the redundancy relationship. Doing so should show how the measure of 'used' information is experimenter-independent. In order to do so, we need to look at the mathematical formulation of redundancy (RED):

$$RED = MI(Feature; PerceptualDecision)+$$

$MI(Feature; MEGVoxelActivity)-$

$MI(Feature; MEGVoxelActivity, PerceptualDecision)$

or schematically, in familiar notation:

$$RED = I(X;Y) + I(X;Z) - I(X;Z,Y) \tag{4.7}$$

This measure bears a few relations to our previously-introduced measure of CMI. For example, if we look at the conditional mutual information between a feature, MEG voxel activity, and a perceptual decision, we see that:

$$CMI = I(X;Y|Z) \tag{4.8}$$

Since it is the case that

$$I(X;Z,Y) = I(X;Y|Z) + I(X;Z) \tag{4.9}$$

We can substitute in order to derive

$$RED = I(X;Y) + I(X;Z) - (I(X;Y|Z) + I(X;Z)) \tag{4.10}$$

Such that

$$RED = I(X;Y) - I(X;Y|Z) \tag{4.11}$$

Leaving us with a demonstration that RED is the mutual information between X and Y minus the conditional mutual information between X and Y given Z. This is just our

interaction information from equation (4.6) above!

Based on our interpretation above, this is a measure of the MI between the feature of the image and the perceptual decision, minus the conditional mutual information between the feature of the image and the perceptual decision, given the MEG voxel activity. This demonstrates that we need not think in terms which refer to our own perception of what is 'redundant'. We can think of RED in terms of *removing* all that information which is *not* contained between all three variables. As we saw in section 4.4.3, CMI provides a measure of the information shared between two variables *over and above* what is shared between three variables. Now we are removing that "over and above" information to get at *just* what is shared between all three variables.[3]

As Zhan et al. describe their methodology: "Our results thus highlight how SIR [Stimulus Information Representation] can be used to investigate the component processes of the brain by considering interactions between three variables (stimulus information, brain activity, behavior), rather than just two, as is the current norm" [Zhan et al., 2019, p319]. This is how they aim to address similar concerns to those of Egan: to look at how information flows from the stimulus to the behaviour, thereby licensing the conclusion that this information is in fact being processed by the system, and is not an artefact of experimentation. Directed specifically at Egan, we might say that this gives a component of the meaning of our talk of constraints imposed by the system itself: downstream areas 'shear off' information and narrow the possible range of content ascriptions.

The results of Zhan et al. are not especially pertinent for our concerns, but we should

[3]See also the relation between redundancy and information transfer over time ("Directed Feature Information") if variables at differing times are compared: [Ince et al., 2017, p1552] and [Ince et al., 2015, p13]. The latter, in particular, treats redundancy measures over time as a relation between CMI measures over time.

note them both for the sake of completeness and because they provide an intuitive understanding of the theoretical framework. The experimenters hypothesise that if a region is involved in supporting decision-making behaviour, we should see a convergence of "diagnostic" feature "representations" (i.e. voxels with RED relative to both the image and the perceptual decision) within that region. Such a region was found in the right fusiform gyrus (previously predicted to support object recognition). Over time, voxels with high RED scores converged on this area, while voxels with low scores diverged and the activity of these voxels decreased in strength through the occipital cortex. So, over time, features which we might interpret as transmitting specific information used in decision-making behaviour are increasingly 'represented' in one specific area of the brain. For further applications of this framework see a review by Sychns et al. (2020) [Schyns et al., 2020].

We should be careful at this point to remember that we are not yet, ourselves, committing to the representational status of these voxel regions, nor are we making concrete suggestions as to the content. Indeed, the authors themselves note their loose representational talk. For example, when discussing the 'nun' features (features which, when we overlay them on the original image, are focused on what we can see to be the faces of the nuns), they remark that "it would be naive to assume that the nun's face is represented as such in any of these regions, but we need a broad view of the information-processing, which this model affords." [Zhan et al., 2019, p324]. What has not been done, and what we still need to do, is to extract an explicit theory of representational content from the theoretical framework employed.

4.5 Objection: over-generalisation?

I have stated that I wish to unlock the implicit theory of content determination in cognitive science. I have argued in this chapter that there is a theoretical framework behind the implicit theory, and that the theoretical framework is information theory. A question I have not engaged with is whether this implicit theory, and its information-theoretic framework, are representative of cognitive science as a whole.

In many ways it does not matter whether every researcher and every paper in cognitive science uses this implicit theory. It does not even matter if some researchers have, consciously, an explicit theory which is nothing like the type I have outlined or indeed is in conflict with it. What matters is whether an implicit theory of content is available from within cognitive science, and whether that implicit theory has the contours outlined. I have set out some restrictions I take to be key for overcoming Egan's concerns, and within those restrictions it looks as though there is an implicit theory and it does have the contours I outline. Only by providing evidence of this view in the form of examples and by (hopefully) providing a sensible reconstruction can I make that case compelling. If there are other views one could in principle extract from other areas of cognitive science, this does not detract from my own project. If cognitive scientists frequently use content talk as a gloss (as may well be the case) this does not mean that within the parameters I have outlined, one cannot extract an implicit theory from other sources within the discipline.

Given this possibility, one has two options. One can either be a pluralist or a revisionist. If one is inclined to be a pluralist, this involves allowing that content talk throughout cognitive science is liable to differ, and that no one view on content should predominate. Even in cases where content is a gloss, we may allow that - in those areas - we

need no theory of content. We might then quibble over word use (what gets to be called 'content'), but the pluralist would see nothing much hanging on such debates. Pluralism might be motivated by noting that different regions of cognitive science pursue different explanatory projects (answering either how- or why-questions, for example), and that some content ascriptions are relevant for some projects, but not for others.

If one is inclined towards revisionism, this involves arguing that just one theory of content should win out. If it is the case that other areas of cognitive science differ in their implicit theory of content, the revisionist might think that we have something better to offer them. If there is no implicit theory of content and it is being used as a gloss, we might think that we could add something to that work by supplementing the gloss with a theory-guided content attribution. Indeed, one might think that progress can be made in cognitive science as a whole with a unified, explicit, theory of content determination. Revisionism might be motivated by noting that some areas of cognitive science are pursing an explanatory project using an unsuitable theory of content.

4.6 Conclusion

This chapter provided an example of the use of information theory in two areas pertinent to content determination: it can model both what the cognitive system picks up from the environment, and what it uses to perform a cognitive task. Information theory, far from being a simplistic account of the correlation between two variables, provides the means to generate incredibly precise content ascriptions constrained both by inputs which cause neurons to respond, *and* downstream processing.

The studies used in this chapter identify the item with which the underlying neural mechanism has maximal mutual information as the content of the neural representation supported by that mechanism. This is what maxMI claims to be the representational content of all neural representations: that item with which the representational state has maximal mutual information. So, the assumption must be shown to generalise, and must be shown to be justified. In the final chapter I argue that maximal mutual information provides the item in the environment about which the system has "available" information (section 7.5). This is what allows content to play a role in proximate explanations.

However, before we can generalise the account, we need to understand how information theory can be used to model the cognitive system more broadly. So, in the next two chapters I discuss the constraints on applying information theory to the cognitive system. In chapter five, I outline how to specify random variables within the cognitive system and in the external environment. This provides us with the formal apparatus to apply information theory to the cognitive system. In chapter six, I introduce functions as a way of overcoming the "reference class problem" - a problem concerning how to limit the range of possible external items relevant for comparison of maximal mutual information values. In Finally, in chapter seven, I outline maxMI in full.

Chapter 5

iRVs and eRVs

5.1 Introduction

In the previous chapter, I argued that, for those regions of cognitive science which address Egan's concerns, the background theoretical framework is Shannon's information theory. Later, in chapter seven, I will spell out the implicit theory of content within those regions of cognitive science in information-theoretic terms.

However, before setting out the implicit theory of content, we need to reflect on the conditions under which information theory can be applied to the cognitive system. This will prove to be crucial, since it will lead us to provide some relatively severe constraints on what we take to be the relevant features of the cognitive system and the external environment for the purposes of a theory of content.

Reflection on the application of information theory is prompted by Shannon's warning, set out in section 5.2. Shannon warns that the application of information theory

must be made carefully, with due consideration to the specific mathematical posits of the theory.

In this chapter, I will highlight a mathematical model central to information theory - the random variable. Random variables form the basic building blocks of information theory. Without specification of what we model as random variables, we are unable to apply Shannon's information theory proper to the cognitive system. However, I will argue that specifying random variables relevant to content determination requires constraining what can serve as both representation and content. We must identify specific outcomes and probabilities over those outcomes.

First, I deal with applying the random variable model to the system. I introduce the term "iRV" to stand for the "internal random variable" which models a representation. In general, random variables require a range of specifiable outcomes, with a probability distribution or density over those outcomes, which collectively sum to unity. I argue that the iRV must be constrained to include only those values which downstream systems are causally sensitive to (sections 5.3 and 5.6) and that the probabilities of the outcomes are determined by the response profile of the physical cell which we apply the model to, in conjunction with the probability distribution of the items in its receptive field (e.g. section 5.3.1).

Second, I consider applying the random variable model to items outside of the representation itself. I introduce the term "eRV" to stand for the "external random variable" which models content. I argue that the eRV must be limited to those outcomes which are detectable by the sensory interfaces of the system (section 5.4). This does not entail that eRVs can only be used to model very basic contents detectable by, say, photorecep-

tors; I show how invariance mechanisms determine the complex discriminatory profiles of downstream neurons, allowing us to model complex external items in terms of the outcome values which can be detected by the system (section 5.4.1).

I end (section 5.6) by considering whether stipulating a random variable relies on pragmatic considerations. Given the degree of flexibility regarding which aspects of the system and environment could, in principle, be modelled as a random variable, I raise the concern that we may be guided by what is pragmatically beneficial. I answer that we specify random variables following constraints provided by those elements of the system which are actually relevant for performing cognitive capacities, rather than our own independent aims and interests.

5.2 Shannon's warning

Following the 1948 publication of Shannon's *A Mathematical Theory of Communication*, information theory became a framework favoured by theorists working in a range of disciplines dealing broadly in signal transmission. This adoption quickly spread to those working in fields with no obvious relation to the original home of information theory, telecommunications. In economics, for example, information theory has been used to model a myriad of phenomena, such as the behaviour of agents with limited access to information (e.g. see [Yang, 2018]). As we have explored in previous chapters, in the case of cognitive neuroscience this trend is still going strong.

Noting the rise in popularity of information theory, in 1956 Shannon sought to quell the enthusiasm which, at times, led to a hasty, unsystematic application of its key con-

cepts. Shannon published a short bulletin, *The Bandwagon*, urging theorists to maintain "a thoroughly scientific attitude". He warned that they should not be enticed by "a few exciting words like *information*, *entropy*, *redundancy*" into thinking that information theory can "solve all our problems" [Shannon, 1956].

Shannon notes that information theory was initially "aimed in a very specific direction", intended as "a technical tool for the communication engineer". Information theory, we are reminded, is "essentially a branch of mathematics, a strictly deductive system". While excited theorists assumed the explanatory potential of information theory across a wide variety of domains, they did not always stop to consider whether this technical mathematical tool, with its roots in telecommunications, was "relevant to such fields as psychology, economics, and other social sciences". Shannon goes on to suggest that if information theory is applicable to these domains, it must be shown experimentally. For example, if we think that neural assemblies transmit information in the information-theoretic sense, we must identify some testable hypotheses in support of the proposal - then test them.

Much progress has been made since the 50s, and many empirical questions are being settled. A substantial body of work exists in the domain of neural information theory to make such predictions and test them in the case of applying information theory to the brain (for an overview, see [Stone, 2018]).

In the domain of informational teleosemantics, however, we not only, in the main, lack testable empirical hypotheses - we lack something much more fundamental; we have no understanding of how the mathematical model of information theory is supposed to relate to the thing we are trying to model - the cognitive system and its way of deter-

mining content. It is not even clear, that is, that there are any properties of the cognitive system which sensibly correspond to properties of the mathematical model. For example, when we discuss the amount of mutual information between some representation and the external environment, what is it, exactly, that we are taking to be modelled by the random variables, what are the relevant probability ranges, what defines the corresponding entropy, where are the channels? and so on. Are we simply getting carried away with exciting terms, hoping information theory can provide a black-box solution to our representational problems?

This is not a problem for theorists who do not use Shannon information proper. Consider Neander's work, often considered to be the leading elaboration of informational teleosemantics, as described in chapter two. According to Neander, there is a need to define our *own* concept of information based on "desiderata [which] must be met by an analysis of the notion of information for particular theoretical purposes" [Neander, 2017b, p145]. For Neander, this leads to an analysis which is much more lightweight than Shannon's mathematical notion, invoking loose historical statistical regularities [Neander, 2017b, p146].

More recently, however, informational teleosemantics has taken on more explicit commitments to Shannon's mathematical framework. Recent theorists such as Marc Artiga [Artiga et al., 2020], Manolo Martinez [Martinez, 2013], and Stephen Mann [Mann, 2018], rely heavily on information theory proper. So does the current proposal, taking our cue from cognitive neuroscience and the background theoretical framework which we investigated in the previous chapter. For us and for other contemporary teleosemanticists working within Shannon's framework, spelling out exactly how information theory is

applicable to the cognitive system is essential if our work is to be set on sure theoretical foundations.

5.3 Internal random variables (iRVs)

What is a random variable?[1] Interestingly, the textbook definition has a form of realism written in. Here is an example: "A random variable X is a function that maps each outcome x of an experiment (e.g. a coin flip) to a number $X(x)$, which is the outcome value of x." [Stone, 2015, p26]. Imagine you are watching a coin being flipped, notebook in hand. Each time you see that heads is showing, you jot down the number '1'. Each time you see that tails is showing, you jot down the number '0'. You have, in effect, created a random variable modelling the coin flip with the definition given by

$$X = \begin{cases} 1, & \text{if the outcome is heads,} \\ 0, & \text{if the outcome is tales.} \end{cases} \tag{5.1}$$

What the random variable models is a coin flip, and is constrained by the properties of flipping coins. Imagine that this coin cannot land on its side. There are, therefore, only two possible states of affairs following a flip: heads or tails. We can label heads x_h and tails x_t. This provides us with our **alphabet** for our random variable: the outcomes which we then map to (typically numerical) values. So, our definition above tells us that for the random variable X the mapping rules are $X(x_h) = 1$ and $X(x_t) = 0$.

We know that, given our observations of the coin flip, each outcome has some given

[1] The following exposition is taken primarily from James V. Stone, *Information Theory: A Tutorial Introduction* [Stone, 2015].

probability of obtaining. This allows us to define the **probability distribution**[2] of the random variable. This tells us how the probabilities of obtaining an outcome are distributed across the outcomes. We write the probability that we will get heads as $p(X = x_h)$ and tails as $p(X = x_t)$. This represents the probability that the outcome of X (the random variable used to model the coin flip) will be either heads or tails[3]. The probability distribution is written as $p(X) = \{(p(X = x_h), p(X = x_t)\}$. This is typically abbreviated to $p(X) = \{(p(x_h), p(x_t)\}$ [Stone, 2015, p24].

This is a fair coin, so on average we get heads half the time and tails half the time. In this case, we have $p(X = x_h) = 0.5$ and $p(X = x_t) = 0.5$. So the probability distribution for X is $p(X) = \{0.5, 0.5\}$.

So, for anything we aim to model as a random variable, we need to identify some outcomes and we need to identify a probability distribution over those outcomes where the probability sums to unity (which can be the result of normalisation). Specifying items to be modelled as random variables is crucial to applying information theory, since the key information-theoretic concept, **entropy**, is determined by the properties of the random variable used to model the entity. The entropy of a random variable is given by the average *surprisal* of each outcome obtaining, which is defined as $1/p(X = x)$.

Mutual information - which, as we saw in the previous chapter, the theoretical framework of content attribution in cognitive science relies on - involves the relation between two random variables: one modelling 'internal' items - the representational states within

[2]This random variable is *discrete* meaning that the values are distinct. For a *continuous* random variable in which each value cannot be separated individually (think of continuous amounts of water filling a jug) the equivalent concept is the probability **density**.

[3]Note that this implicitly relies on the fact that heads and tails exhaust the possible options. Formally, it is assumed that $p(X) = 1$ (the sum of all marginal probabilities is 1).

the system - one modelling 'external' items - the items in the world which serve as content.

So, we can now refine our initial realist objective; we have to find something which is modelled by a random variable on the side of the cognitive system, and a corresponding random variable on the side of the external environment. I will call, for short, the former an **iRV** for 'internal random variable' and the latter **eRV** for 'external random variable'.

5.3.1 Initial outcome ranges and probability distributions

Where to begin? In the previous chapter I argued that the best place to look for an implicit theory of content is in those studies which deal with the interface between the organism and the environment. This is due to the fact that these experimenters are explicitly interested in specifying the precise environmental features which drive an internal response, so external items feature explicitly in their explanations. When attempting to provide the features of the system which map onto the mathematical model, we would do well to continue this trend for a few reasons.

First, we will begin our exposition with a relatively simple example. Early perceptual systems very quickly increase in complexity, so starting at the most basic level (which is far from basic) will help us get a grip on how the mathematical model applies.

Second, the studies in psychophysics and cognitive neuroscience we have been looking at are concerned with this basic relation, so our general methodology encourages us to follow them.

Third, if we find we *cannot* show how the model applies in the most basic case, we will probably struggle with more complex cases, so it is best to start here as a proof of

concept.

Photoreceptors and transduction

I will begin with a brief overview of how photoreceptors work[4]. I will then attempt to show how photoreceptors can be modelled as random variables, and try to present a plausible suggestion for corresponding environmental random variables. I will go into a little of the biological detail. This is important to get an idea of the specifics which the random variable model generalises over. We need to provide some details in order to achieve the model-to-system mapping required for a realist interpretation.

Photoreceptors are cells in the retina which respond to light. As is well known, the retina contains two types of photoreceptor, rods and cones. Rods respond to monochromatic features of light, while cones are responsive to differences in light which determine colour. We will focus on rods for simplicity. There are about 120 million rods in the retina. At the back of the cell, embedded in the retinal wall, they each contain "thin membrane plates" known as "*lamellae*" [Tovée, 2008, p29]. In rods, the lamellae are individual disc-like structures suspended in the rod. Bound to the lamellae are "photopigment molecules" known as "*rhodopsin*". Rhodopsin is constituted by opsin and retinal. Retinal can exist in the "straight" chain form (all-trans retinal) or the "bent" form (11-cis retinal). Only the 11-cis form of retinal can bind to opsin. When a photon hits the lamellae, the 11-cis form of retinal is caused to transform into the all-trans form, which can no longer bind to the opsin. So, the rhodopsin breaks down and the lamellae is "bleached".

When this happens, the rod becomes *hyperpolarised*. Unlike many cortical neurons,

[4]The exposition here is mostly taken from Martin J. Tovèe's *An Introduction to the Visual System* [Tovée, 2008].

photoreceptors, when inactive, have open ion channels. The channels are kept open by a chemical called cGMP[5]. The breakdown of rhodopsin causes levels of cGMP to drop, which in turn leads to the closure of the ion channels within the cell. This creates an increase in the resistance between the inside and outside of the cell, lowering the cell membrane potential (measured in millivolts, mV). This ultimately leads to a reduction in the amount of neurotransmitter (glutamate) sent to the post-synapse ganglion cells as the decrease in voltage propagates along the axon of the photoreceptor.

Unlike cortical neurons, photoreceptors do not emit action potentials - which are all-or-nothing - they emit *graded* potentials [Purves et al., 2001]. We can think of the reduction in glutamate resulting from hyperpolarisation as continuous. The degree of hyperpolarisation depends on (ignoring the effect of noise due to heat) the number of photons absorbed. There is an upper limit to the degree to which the the voltage can decrease, and there are upper and lower bounds on the intensity levels of photons that particular photoreceptors are sensitive to. Sensitivity to intensity is unlikely to be linear, since an S-shaped response profile is much more efficient (see [Tovée, 2008, p37]). Greater responsiveness is to be expected at the middle range of intensity. In other words, changes in the number of photons absorbed will lead to greater changes in membrane potential at the middle of the intensity range than at the extremes. In information-theoretic terms: for bounded random variables, which the photoreceptor can be modelled as, maximum efficiency of information transfer is achieved when the probability density is uniform, which the S-shaped sensitivity profile ensures, provided Gaussian inputs.

What does all this mean for modelling the rods as random variables? Recall that

[5]Cytoplasmic cyclic guanosine 30–50-monophosphate, [Tovée, 2008, p30]

we need some aspect of the modelled that corresponds to its 'outcomes' which, in the model, are its **values**. At this point we should also note that the rods should be modelled as continuous random variables. This is because whatever we model as the outcomes - the decrease in voltage, the decrease in glutamate production or whatever - they will be continuous due to the graded nature of the cell's response. So, we need to identify the **probability density function** of the outcomes of the cell. More on this below.

5.3.2 Subsequent outcome ranges

We have investigated how to model the initial interface as a random variable. In this section I provide a little detail about how we can stipulate an iRV for slightly further downstream neurons. Information theory is not just a useful tool for measuring initial sensory systems. Provided we can identify outcomes and a probability distribution over those outcomes, we can stipulate an iRV which is relevant to content determination. In this section, I use ganglion cells and simple cells as examples demonstrating how to extend the iRV model generally.

Ganglion cells

Photoreceptors connect to retinal ganglion cells. A large number of photoreceptors connect to each ganglion cell (at a ratio of about 126:1 on average). The photoreceptors which connect to a given ganglion cell define that cell's **receptive field**. In general, moving up the visual hierarchy, the receptive fields of post-synaptic cells tend to increase, incorporating a greater number of inputs.

Ganglion cells do not respond on the basis of a simple threshold of active photorecep-

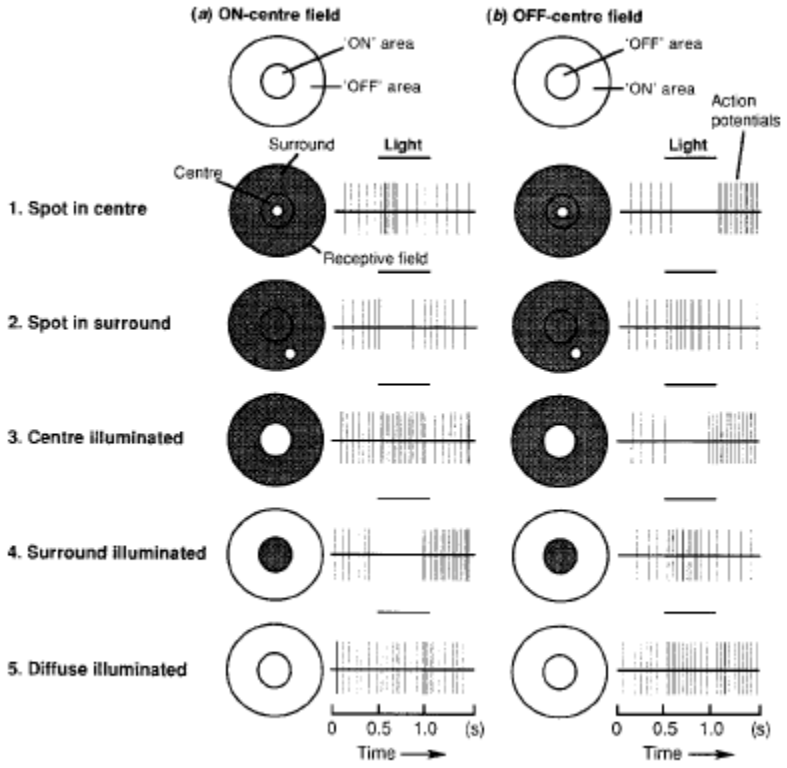


Figure 5.1: Reprinted from Tovèe (2008)

tors in the cell's receptive field. Instead, photoreceptors are structured in what is known as a "centre-surround" organisation [Tovée, 2008, p34].

Photoreceptors are arranged in concentric circular patterns within the retina. Ganglion cells are connected to circles of photoreceptors with a centre and a surround circle (as in figure 5.1). Each ganglion cell is preferentially activated with either an ON-centre, OFF-surround activation pattern (ON-field), or an OFF-centre, ON-surround activation pattern (OFF-field). For ON-field ganglion cells, optimal responsivity of the cell is achieved when all photoreceptors in the centre of the circular arrangement are active

(i.e. hyperpolarised) and all photoreceptors in the surround are inactive. Conversely, OFF-field ganglion cells are optimally activated when the surround photoreceptors are all active, and the centre photoreceptors are inactive.

This is known as an **opponent channel** since the centre and surround are in 'opposition' - for example, in ON-field ganglion cells the centre of the receptive field activates the cell, while the surround inhibits the cell. So, if both the centre and surround are ON, the activation field will cancel out and the ganglion cell will default to its average firing rate.

Ganglion cells, unlike photoreceptors, emit action potentials rather than graded potentials. An action potential "is a rapid sequence of changes in the voltage across a membrane" [Grider et al., 2019]. This rapid change in voltage is propagated down the cell's axon. When an action potential is emitted, a sudden, sharp change in voltage is transmitted.

Since action potentials do not vary in voltage, voltage level cannot be what we model as our outcome value for the purpose of modelling the random variable. More precisely, we *could* model the voltage level as a random variable with a single outcome and a probability of recording that voltage (with the relevant comparison set being other possible voltage levels, as in the photoreceptor case) at unity. However, this would not allow us to model the content-determining-relevant aspect of the cognitive system. Since there is no change in voltage, the post-synaptic cell has no way of differentially responding to differing voltage levels - the post-synaptic cell is insensitive to the voltage level as a means of changing its response profile.

As is well known, the relevant aspect of the ganglion cell's output which the post-

synaptic cell responds to is a *temporal* characteristic of the ganglion cell's action potentials. There are typically two temporal profiles which are thought to be relevant[6] and which define two 'codes': the rate code and the timing code.

The rate code essentially involves the mean firing rate of the neuron. If the post-synaptic cell is responsive to the rate code, it will demonstrate differential activity based on how rapidly the cell is firing. A cell may send out an action potential at either a low or high rate measured in spikes per second. A spike refers to the rapid change in voltage.

The timing code is more complex, and involves the position of each spike. Imagine that a neuron fires rapidly for a second, then slowly for a second, then rapidly for a second. Imagine that another neuron fires rapidly for two seconds, then slowly for a second. The mean firing rate of the two neurons may be the same, but the post-synaptic neuron will respond differently to each cell, since the patterns of firing rates vary.

There is ongoing debate about the type of code used by the ganglion cell's post-synaptic cells. Often, considerations of information transmission efficiency are used to adjudicate between different hypotheses [Van Rullen and Thorpe, 2001]. The empirical facts change the aspect of the cell which we model as outcome values. However, at the most general level, we can say that the outcomes will be given by the relevant temporal profile of the ganglion cell's action potentials. In the simplest case, with the rate code, our *highly idealised*[7] random variable would look something like:

[6] Although there are many more (infinite?) logically possible profiles which post-synaptic cells *could* be responsive to. See Gallistel (2017) [Gallistel, 2017] for an investigation into some alternative candidates for the output of a cell.

[7] In reality, we would see much larger spikes per second, since action potentials are emitted on a very small time scale. It is also not clear if the post-synaptic cell is responsive in either a continuous or discrete way to changes in firing rate. For simplicity, I have chosen discrete.

$$X = \begin{cases} 0, & \text{if the outcome is 1 spike/s,} \\ 1, & \text{if the outcome is 2 spikes/s,} \\ 2, & \text{if the outcome is 3 spikes/s.} \end{cases} \tag{5.2}$$

To reiterate, the key point here is that what outcome we use in our model is dependent on what the post-synaptic cell is sensitive to, specific empirical details aside.

The probability distribution of the firing rate of the cell will be determined by two things. First, the receptive field of the cell and its status as an ON-field or OFF-field cell. Second, by a corresponding function of the the probability ranges of the photoreceptors in the ganglion cell's receptive field. As we know, these are determined by the probability of photons hitting the lamellae and the photoreceptor's response profile. So, ultimately, the probabilities of some distribution of photons hitting the lamellae contribute to the determination of the probability range of the ganglion cell. However, unlike photoreceptors, it is not the probability of a photon hitting the lamellae which determines the probability distribution of the ganglion cell - it is a particular *distribution* of photons hitting *several* lamellae together which determines the probability range.

Simple cells

Neuroscientists mark a distinction between 'simple' and 'complex' neurons. Simple neurons are defined as those which have OFF and ON regions, and for which there is a "push-pull" response profile; the cells either hyperpolarise or depolarise depending on the 'contrast' of the input stimuli (the value of the metric they are responsive to) [Martinez et al., 2005]. In the visual system, simple neurons take input directly from the lateral geniculate nu-

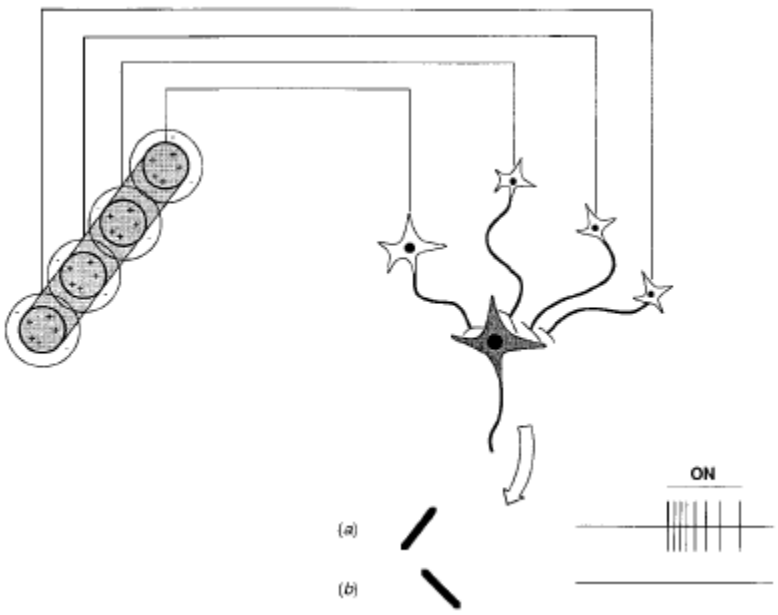


Figure 5.2: How orientation selectivity is thought to build from opponent channels.

cleus (LGN). Complex neurons take input from simple neurons.

Consider the so-called orientation selective cell, as seen in figure 5.2. This is a cell which is hypothesised to respond to bars of varying orientation within the visual field. The responsivity of the orientation selective cell is a product of outputs from multiple input ganglion cells in its receptive field, via the LGN. Specifically, optimal response is found when the stimulus is positioned such that its projection on the retina fits within the inner circle of the centre-surround structure of the photoreceptors within the receptive field of the ganglion cells. Results of a classic study by Hubel and Wiesel show that the response of the orientation cell appears to diminish when the stimulus deviates from a given orientation in either direction [Hubel and Wiesel, 1962]. In addition, the orientation selective cell appears to be inhibited by activity at the edges of its inputs, suggesting

that the optimum response is found by a stimulus with a specific, finite length.

The application of the random variable model should be familiar by now. We have pointed out the various inputs which define the responsivity of the cell, and the application of the model relies on taking the output of the cell which the post-synaptic neuron is responsive to as the outcome. This depends on things like the 'code' used by the cell, which we understand as that aspect of the cell's firing which the post-synaptic neuron is sensitive to. If we assume a rate code, the random variable will be as in equation 5.2.

As before, the probability range will be determined by both the features of the target cell and the features of its pre-synaptic inputs. Ultimately, this will depend on the probabilities of an increasingly complex pattern of photons on the surface of the retina.

5.4 External random variables (eRVs)

In principle, there are a vast number of ways to model any item as a random variable. What we need is a principled way to model external items relevant for determining content. In this section, I expand on Neander's insight[8] that the external items relevant for content are those which are constrained by the discriminatory capacities of the organism.

As Neander puts it, each external item has a "determinate" value along a range of possible values known as "determinables" [Neander, 2017b, p128]. For example, the distance between eyes varies across individuals. We can describe the specific distance between the eyes of a specific person as a determinate value relative to the determinable range of possible distances. Neander suggests that each determinable range corresponds to an internal range: for each "range of values C_1 ... C_n of an environmental determinable C"

[8] As discussed in chapter two section 2.4.1.

there is "a range of values R_1 ... R_n of an inner determinable R" [Neander, 2017b, p128]. A range of possible distances between eyes is, according to Neander, associated with a range of possible states of an internal representation.

I argue that we should view this as a constraint on what we specify as the outcomes of the external item. In order to specify an eRV a change in which results in a change in the system itself, the system must be causally responsive to the values of the eRV. The best, perhaps only, way to ensure this is to include in our eRV model a range of outcomes C_1 to C_n *only if* there is a range of values R_1 to R_n which are causally sensitive to the values of C. In short, the values of the iRV used to model R determine the values of the eRV used to model C.

Consider the orientation selective cell. Let us assume that it is the function of the orientation selective cell to detect lines. This provides us with a space of entities to model as external random variables - i.e. anything which is a line[9]. However, how do we apply the random variable model? What are the parameters within which we vary the lines in order to take down outcomes and attribute probability ranges?

To take the most basic case: for photoreceptors, we saw that the probability range of their outcome values are determined by the tuning curve of the photoreceptor (its sensitivity profile) and the probability distribution of the obtaining of photons on the surface of the cell's lamellae. This relation between the probabilities gives us an obvious constraint on the elements to be modelled as outcomes in the eRV in this simple case: the presence or absence of a given number of photons on the surface of the lamellae. It is the probability range of the values of the eRV so construed which determines (via either

[9]See chapter six for a full explanation of how functions enable this.

phylogeny or ontogeny) the response profile of the photoreceptor.

Moving further downstream, orientation selective cells, as we saw in section 5.3.2, have a centre-surround structure built on the centre-surround structure of ganglion cells in the receptive field of the orientation-selective cell. Activation of ganglion cells in the surround of the orientation cell inhibits responsivity of the orientation cell, meaning that orientation cells respond optimally to lines of a certain length. This provides one parameter (with an upper and lower bound) which we can model as an outcome: line length. However, we need only consider those line lengths which the cell is responsive to: line lengths above the sensitivity threshold of the cell are irrelevant for content determination. The cell contains no information which could differentiate between those lengths, so we do not include them in our eRV since this would create an eRV a change in which *does not* result in a change for the system itself.

5.4.1 Scaling: determining eRVs for complex items

If eRVs are to be determined by constraints on discrimination, does the model only work for relatively non-complex cells with straightforward environmental inputs, such as photoreceptors, ganglion cells and simple cells? I argue that we can apply discrimination constraints throughout the system. This is because downstream areas are connected to upstream areas via **perceptual invariance** mechanisms. Initial iRV values determine complex eRV values in virtue of the corresponding complex iRVs being connected to initial sensory surfaces.

Cohen provides the "textbook definition" of perceptual invariance as "stability in perceptual response across a range of perceptual conditions" [Cohen, 2015, p2]. In neural

terms, invariance mechanisms underlie the invariant response of a neuron to variation in the inputs to that neuron. The (in)famous grandmother cell, a hypothetical cell which responds only to one's grandmother, is a prototypical example. Grandmother cells are hypothesised to exhibit extremely sparse responsivity - the cells rarely respond to input, only doing so under highly specific conditions (e.g. the conditions which are present only when one's grandmother is present). There may really be such cells; Quiroga et al. report that, in the hippocampus of one individual, a cell was found which responded to all and only Halle Berry input, invariant across different pictorial presentations. It even responded, it seems, to the written name 'Halle Berry' [Quiroga et al., 2005].

Grandmother cell examples are an extreme case of invariance. However, even cells in early sensory systems exhibit invariance. For an intuitive idea, consider the vast variability of retinal activation caused by any given object, no matter how apparently basic. Light from a smooth rock can be projected to any position on the retina, it can be viewed from multiple angles, under various lighting conditions, through various media of varying viscosity. In order to allow us to recognise the rock across all these conditions, the brain must take what DiCarlo et al. call the high-dimensional "object manifold" [DiCarlo et al., 2012] (the profile of the object encompassing all possible retinal activation patterns) and produce a single, invariant response.

Invariance mechanisms build sparsity of response by performing operations over several inputs. Downstream cells preferentially respond to patterns of earlier responses. Here is a toy example to highlight the idea. Imagine two cells, A and B. Cell A has possible response states A_1, A_2, A_3 (or 'off'). Cell B has possible response states B_1, B_2, B_3 (or 'off'). We can imagine a third cell, C, which responds only when A is in A_2 and B is in

B_1 or when A is in A_1 and B is in B_3. Otherwise, C is 'off'. C's response profile evinces invariance: C is stable across a range of conditions, and does not respond to another range of conditions in the same inputs. If we assume that the combinations C responds to are rare, but A or B being in any one of their respective states is common, C's response profile is sparse relative to its inputs.

Invariance mechanisms enable the system to detect such complex items as the incline of a plane. Degree of incline is calculated via the pattern on the surface of the plane. Statistically speaking, surface patterns tend to be uniformly distributed. Imagine a surface made up of uniformly distributed circles. If the circles are increasingly elliptical and densely distributed towards one end, the system interprets that as an inclined plane. To do this, the system must *also* be able to produce an invariant response to the texture - to build a complex response profile which does not vary according to density of distribution or eccentricity of the texture-relevant shapes [Burge, 2010, pp355-359]. It then compares this representation to the skewered input and produces a representation of an inclined plane.

Invariance mechanisms scale to enable increasingly complex capacities. Object recognition - typically operationalised as the ability of participants to apply labels in expected ways to stimuli - is thought to be facilitated by invariance. Leading classic and contemporary theories of object recognition both rely on invariance as a mechanism. The classic theory, due to theorists such as Tanaka (e.g. [Tanaka, 1997]) and Stryker (e.g. [Stryker, 1992]), suggests that there is a "visual alphabet" which composes to form complex representations of visual objects. In these theories, the building blocks of object recognition are early visual representations of shapes, patterns, colours and so on. Neu-

ral implementations of these representations are thought to be built from cells such as the orientation-selective cells we have been considering. These building blocks are processed together in order to produce complex representations.

The building blocks themselves are developed from combinations of inputs from simple cells. In the same way that simple cell response is built from insensitivity to aspects of the input, and sparse responsivity to other aspects (e.g. to activation of the centres of the receptive fields of the input ganglion cells), object recognition involves sparse responsivity to a collection of the neural inputs comprising the building blocks, but insensitivity to the isolated activation of any one of them.

For contemporary theorists DiCarlo et al., the "'invariance problem' is *the* computational crux of recognition" [DiCarlo et al., 2012, p17]. Invariance, as they understand it, involves producing a single neuronal response to each object category given an extremely high-dimensional retinal input. Sparse responsivity must be developed in such a way that objects which have similar retinal profiles can be neatly distinguished by the system. Elements of the input profile must be grouped, with downstream systems demonstrating insensitivity to isolated members of those groups.

The key claim I am making is this: invariance mechanisms within the system provide parameters according to which complex external items are to be modelled as random variables. The system essentially chunks environmental items in such a way that allows the **parameterisation** of the input. The response of downstream neurons is constrained by highly disjunctive upstream discrimination profiles, allowing us to discover which aspects of complex external items drive a system-side response. These aspects are candidate outcomes for modelling an eRV.

5.4.2 Limits of invariance

However, even once we parameterise the input, there is no guaranteeing which of the many possible related eRVs model the content of the target neural representation.

The number of photons present on the surface of the lamellae, lines in the environment, and tilted planes, each correspond to other items which can also in principle be modelled as external-side environmental variables. For example, the number of photons present on the surface of the lamellae typically corresponds to the light intensity in the vicinity of the receptor. Given the constraints imposed by the number of photons present, we can model light intensity as a random variable. Furthermore, light intensity level will correspond to more distal environmental variables which cause fluctuations in light intensity. The presence of lines will correspond to edges of objects, inclined planes will correspond to depth and size and so on, which can be modelled using random variables provided the values of each map to the values of each lower-level eRV.

As such, we have a potentially unbounded range of eRVs, each of which *prima facie* could be the content of a representation, since mappings from the values of higher-level eRVs to lower-levels are cheap.

In the next chapter I argue that we must limit the range of possible eRVs which could serve as content by way of the **function** of the subsystem containing the neural representation. This will limit the possible eRVs which could be content to those which can, in principle, provide an explanation for the cognitive capacity under investigation. In the final chapter, I argue that we pick representational content from among those eRVs within the set, delineated by both discriminatory capacities and system function, according to which eRV maximises mutual information with a target iRV.

5.5 Information theory requires the brain

In the above sections, I have afforded a significant role to the brain within what I have been calling the cognitive system. One may object that there must be other ways to apply information theory to the system. Surely, there are other aspects of the system which could provide the relevant outcome ranges, and over which we can define a probability distribution.

For example, we might think that we can take a state defined using more recognisably psychological terminology as an output. For example, perhaps we could consider entertaining one concept rather than another. We may take a particular concept as an outcome, and define the probability range over the potential concepts I *could* have entertained.

There are three significant problems with defining outcomes in these terms. First, what I call the **relevance issue**. We need to know the relevant set of psychologically identified entities over which we could define the probability range (which sum to unity). However, concepts have myriad possible connections to each other, each of which could define alternatives one may have chosen. Cars might be contrasted with other possible vehicles, but they may also be contrasted to things with wheels, things with engines, and so on. The choice will be highly contextually constrained, as well as dependent on individual differences in what one *takes* to be relevant. When calculating probabilities for random variables one must define a "reference class" of other outcomes with probabilities which collectively sum to unity (I discuss the reference class problem for maxMI in detail in section 6.2). Given the huge variability, flexibility, and individual differences in the range of possible concepts one *could* have entertained, it seems incredibly difficult to solve the reference class problem for concepts. So, specifying a random variable seems,

at this stage of understanding at least, hopeless.

Second, what I call the **probability issue**. We need to calculate the probability of a given concept being entertained in general (notwithstanding the relevance issue). However, we are infinitely creative creatures, with no obvious *a priori* constraints on our ability to apply concepts to anything. There will of course be situations in which the tokening of a given concept is more or less likely, but individual differences in application, and the multitude of factors which go into making concept ascriptions, make any precise calculation of probability currently intractable.

Third, what I call the **scaling issue**. If we start with concepts, how could we scale *down* to more basic states? This is the inverse of the scaling-up problem usually attributed to the informational teleosemantic approach (see section 8.3). Could we apply the same theory of content to states which cannot be picked out using a psychological framework? It might be that our theory stipulates that representations do not occur at such a level. This is a difficult position to maintain within the current explanatory framework: representation appears, as we have extensively covered, to be explanatory even in contexts where more traditional psychological models, such as concepts, do not apply.

These issues are solved by invoking the brain. The set of relevant alternatives is clear: the alternative possible firing rates or voltage levels output by the cell. How to determine the probability has been the topic of this chapter: distal probabilities, along with response profiles of pre-synaptic cells and the response profile of the target cell define the probability of a given outcome. Likewise with scaling down: we have applied our model to photoreceptors, and it does not get much more basic than that.

Instead, we should afford the brain a central role in content determination. Indeed,

reflection on the requirement of the systematic application of information theory to the cognitive system suggests a novel way of interpreting the relation between the brain and the mind; the brain provides the mind with a quantity of information, the mind takes that information and produces representations with content. The mind here is understood as the totality of the functional relations between brain regions. Without a function, the information amount provided by a neuronal response fails to be 'about' anything: there is no distal content which it specifies. There is only a bare quantity of information. Once we add in the functions, the mind, we have something which consumes information to make distal representations which serve its various needs. Together, the brain and the mind comprise the whole of the cognitive system.

So, I argue that the relevant part of cognitive science, for our purposes, will be branches of **neuroscience**. Specifically, that part which deals with cognitive functions: cognitive neuroscience. If we want a systematic theory of content with foundations in information theory, we must turn to those areas of science which invoke the brain.

5.6 Problematic pragmatism?

I end with a worry about pragmatism. For the purposes of stipulating a random variable, what is to be considered the outcome is a matter of *choice*. As we saw in section 5.3.1, even in our simplified story, there are many elements in the causal chain which result in post-synaptic changes for the ganglion cell. There is an increase in the all-trans form of retinal, a decrease of cGMP, a closure of ion channels (about 2% of a cell's open channels per photon absorbed), an increase in resistance over the cell membrane, a decrease in

membrane potential propagating along the axon, and a reduction in neurotransmitter (glutamate) production. All of these aspects of what the cell does could, in principle, be modelled as the cell's outcome.

It seems like we are plunged into Egan's worry, since the application of the model now looks to be sensitive to pragmatic concerns. Which aspect we choose to consider the outcome, it might be thought, is going to be sensitive to our own interests.

This is the point at which we should draw a vital distinction. I will call any form of choice resulting from maximising various helpful heuristic values - such as easy communicability of the theory, or emphasising a link to pretheoretic interests - **problematic pragmatism**. This variety of pragmatic choice is constrained only by the reception of the theory - how the theory will be received by the public, or how it is intuitively grasped by experimenters.

Another form of pragmatism I will call **unproblematic pragmatism**. This variety of pragmatic choice is constrained only by the initial orientation of the theory; it is an inherent part of theory construction with essentially the same status as hypothesis construction. In this sense, 'choosing' which elements of the system to model as outcome values, for example, amounts to a hypothesis about which elements of the system are relevant to the phenomenon under investigation. In other words, it is a hypothesis about which part of the system does *in fact* perform the role of the outcome within the system itself.

We may be wrong in our choice: we might decide to model the reduction in glutamate as the outcome value. It might turn out that, in doing so, we miscalculate the amount of information available to downstream systems. If realism is indeed right, then changes in

the amount of information available to downstream systems will make a huge amount of difference. It will impact, for example, discrimination abilities (see e.g. [Stone, 2018, p38] for an example). Or, it might not matter whether we model the reduction in glutamate or the reduction in voltage as the outcome. In which case, there will be some story to tell about how these two things are so tightly related as to make no difference to the system itself - they both serve equally well as the outcome *in reality* since they *actually* have a very tight relation. Either way, there will always be a corresponding story to tell about the modelled if the model is wrong (or indeed if it is right). We might have failed to correctly map the model to the system, but it does not follow that there is *no* correct mapping of the model to the system.

Whatever the case, unproblematic pragmatism results from a choice which is constrained by the features of the system itself. If I am right, and choosing which element in the causal chain of the photoreceptor is an *un*problematic pragmatic choice, we evade Egan's worry and can maintain a form of realism. Realism requires pointing to an element of the modelled and providing a justification for placing that element of the modelled in correspondence with an element of the model. This just means that we need to provide a *justification*, constrained by the features of the system itself, for which elements of the system we place in correspondence.

Which element of some internal component should be modelled as the outcome value is an empirical question; it is that aspect of the cell which differentially[10] drives the response of the post-synaptic neuron. Let us turn to existing practice. James Stone writes that it is the *voltage* differential of the photoreceptor which tends to be modelled as the

[10] i.e. if we apply Mill's method of difference.

outcome, but that "transmitter quantity and voltage are related by some (simple?) function (probably a sigmoid), so I would think this is not a vital detail unless you are modelling a very fine granularity" (personal correspondence).

How can we justify this last comment while avoiding pragmatism? Well, as we mentioned above, since there is a very tight relation between the two quantities, there will likely (an empirical matter) be no difference at the level of the system itself *in its content-determining capacity*[11] regardless of which choice is made. The constraints at the level of the system under-determine the choice at the level of the model. So, the outcomes may either be the decrease in voltage levels or the decrease in glutamate release.

So, an author such as Egan may say the following: the under-determination of the model relative to the system it models shows that the only choice of outcome left is one which is problematically pragmatic. We decide between voltage levels and glutamate levels on the basis of a choice totally unconstrained by the system itself, which suggests the choice is constrained only by heuristic values.

I suggest the following line of response: we do not have to choose between the two. One of the benefits of using a model is that we can generalise over specifics. In this case, the outcome can just be considered as the voltage-transmitter subsystem of the photoreceptor. If it is true that there is no difference between the two at a grain detectable to downstream systems (again, I must stress, an empirical hypothesis which *may* be incorrect) we simply generalise over the whole subsystem which includes the voltage change and the inhibition of glutamate as the outcome value. When it comes to taking an outcome to convert into our variable value, we can, for practical purposes, choose from any

[11] There may be some difference at a level which is not registered by the system in all the aspects of that system which we explain by appeal to content.

measurable quantity from this overall subsystem. We can, for example, just take a voltage reading. This looks like a problematically pragmatic choice of outcomes, but this is just an ordinary pragmatic choice of what reading we take from within the outcome system. It is like using either fMRI or EEG.

So, for example, the random variable would look something like this (note that the values are continuous, as are the outcomes):

$$X = \begin{cases} 0, & \text{if the outcome is -40mV,} \\ \ldots & \\ 0.01, & \text{if the outcome is -41mV,} \\ \ldots & \\ 1, & \text{if the outcome is -60mV.} \end{cases} \tag{5.3}$$

For theoretical purposes, the outcome maps onto the voltage/transmitter inhibition subsystem. For practical purposes, we take whatever sensible reading we can from this system.

The probability density function is, thankfully, significantly easier to determine. This will simply be given by the probability of getting our outcome. Again, for practical purposes, we may wish to measure the probability of the voltage. Since there is (probably) a straightforward function between voltage level and glutamate inhibition, we could very easily translate this into probability of a given reduction in glutamate. The probability will be determined by two things: the response profile of the photoreceptor and the probability of a photon hitting the receptor. As we mentioned above, the response profile

of the photoreceptor will likely be a non-linear cumulative distribution function of the Gaussian probability distribution of photons hitting the receptor. If this is true, then there should be, more or less, an equal density of probabilities across all values. The empirical details may vary, but the probability density function will always be determined by these factors.

5.7 Conclusion

In this chapter, we investigated how to provide a non-pragmatic application of information theory to the cognitive system.

I argued that those elements of the system which we should model as iRVs are constrained by those elements of the system which downstream systems are causally responsive to. I aimed to show how, from very early sensory receptors, we can stipulate increasingly complex iRVs throughout the cognitive system.

I argued that external random variables are constrained by parameters given to them by the invariance mechanisms of the system. Essentially, the system parameterises its inputs. Constraining external items in this way allows the application of the random variable model in such a way that the modelled is relevant to the system itself. In general terms, this analysis suggests that high-level representational states can be given in terms of compositional arrangements of low-level states: low-level iRVs plus rules of composition given by invariance mechanisms define complex iRVs, and allow the modelling of complex external items as eRVs, given their relation to the low-level iRVs.

The above emphasises that content must be given in terms of the *internal connectivity*

of the system. We need to know both what the upstream inputs are which determine an iRV, and what downstream systems are responsive to in order to model elements of the system and environment as outcome values.

Finally, I argued that the constraints imposed by carefully applying information theory to the cognitive system, at least given our current level of understanding of psychological states, confine our search for an implicit theory of content to cognitive neuroscience.

This chapter, hopefully, sets out how to apply information theory in a systematic way to the cognitive system and external environment. Random variable specification provides us with the minimal requirements of the formal model of information theory. However, maxMI also involves a comparison of mutual information values of multiple eRVs for a given iRV. In the next chapter, I outline a concern about applying maxMI in this way - the "reference class problem" - and suggest that focusing on functions can help us overcome the concern.

Chapter 6

Functions

6.1 Introduction

In the previous chapter we explored how to stipulate random variables for the cognitive system. Specifically, we looked at the outcomes to be modelled as values and how to model probabilities across those values, with an eye to doing so constrained by the requirements of a theory of content. Stipulating random variables is necessary for applying information theory to the cognitive system. This is crucial for maxMI, according to which the content of a representation is the item with which the representation has maximal mutual information.

I argued that values of an eRV should be limited to those outcomes which can be discriminated by sensory interfaces of the cognitive system. However, this limitation leaves a wide range of possible eRVs which any given representation may be related to. Lots of things a system can detect might be relevant to performing a given cognitive

capacity. For example, some subsystem may be connected to upstream areas capable of discriminating both eyes and ears, but only one may actually be used for face recognition (see the toy example in 1.4.1). Indeed, there may be innumerable things which the system could in principle discriminate, but which play no role whatsoever in enabling some particular cognitive capacity. So how do we determine the relevant set of eRVs?

This is a version of the reference class problem, explored in section 6.2. This problem concerns the fact that maximal mutual information is determined by the range of eRVs which we are comparing against - the reference class of eRVs related to the iRV.

I answer (section 6.3) that the relevant class of eRVs is given by the function of the subsystem containing the target representation. This limits the eRVs to those which are, in principle, able to be used by the system to perform a given cognitive capacity.

I consider two leading theories of function, proposed by Larry Wright and Robert Cummins, respectively (section 6.3.1). Wright functions are etiological and concern the evolutionary or learning history of the organism, while Cummins functions are non-etiological, concerning only what the system currently does.

I argue that the relevant notion of function used to delimit the eRV set is Cummins'. I call these C-functions, after Cummins. The argument consists in arguing that functional ascriptions in cognitive neuroscience - which is the region of cognitive science we limited our search to in the previous chapter (section 5.5) - are made using C-functions. This is because functional ascriptions are made even in cases of extreme pluripotency - the highly flexible transmission of information across the cognitive system. I then outline in some detail how C-functions provide the requisite delimitation of the eRV set (section 6.5).

I then consider some objections to the use of C-functions for a theory of content. Primarily, I consider Neander's objection that C-functions do not allow malfunction or dysfunction (section 6.6.2). I argue that C-functions must allow for dysfunction given the prevailing research paradigm of using dysfunctional patients to infer facts about well-functioning individuals. However, I argue that if we type-identify components of a system with respect to their phenotypic structure, we can assess what that same component contributes to the capacities of another system. If a component in the system under investigation does not enable the same capacity as the same type of component does in another system, we can conclude that the component under investigation is dysfunctional, even on Cummins' account.

I end with a brief discussion of swampman cases (section 6.7). I argue that while swampman should be a concern for teleosemantic theories aimed at providing proximal explanations, the use of C-functions avoids the concern in our case.

6.2 The reference class problem

Applying information theory to the cognitive system requires us to address what is known as the "reference class problem" (for an overview, see [Hájek, 2007]). The reference class problem is that there are an infinite number of outcomes against which we could, in theory, compare any given outcome when attempting to determine conditional probabilities involving that outcome, and which outcomes we compare against will alter the conditional probability measure.

As Millikan points out, the reference class problem is an issue for informational the-

ories of content. Her description of the problem is as follows:

> It could of course be that given all space-time as the reference class, the boiling of water does raise the probability that it is at 212 °F. But boiling is likely to raise the probability more that water is one or another of various other temperatures, there being no reason to think that one Earth-atmosphere is an especially common pressure. Similarly, we have no evidence that an elephant-like head at one end raises the probability of an elephant-like tail at the other throughout all space-time, and it is quite certain that the direction of the North Star does not raise the probability of that direction being geographic north universe-wide. [Millikan, 2013, p136]

The basic point is that when we calculate conditional probabilities, we implicitly limit the range of comparisons. For our purposes, consider neural representations: the firing of a neuron is compared against various external states obtaining in order to calculate conditional probability values. But we cannot possibly compare against *everything* in the universe. Even if we could, we would find that the measure we make varies with the 'reference class' - everything else we deide to use for the comparison.

So, our question is whether there is a way to provide a reference class which fixes the comparisons we make while not being arbitrarily or pragmatically chosen.

In the last chapter, we essentially proposed a way to solve this for the *values* of external items related to neural representations: we limited the eRV outcome values to those which can be detected by sensory receptors within the receptive field of the relevant neural population. This, indeed, is similar to Millikan's proposed solution: "I suggest that

the only relevant or non-arbitrary reference class is the class of (candidate) signs that [...] the animal must be able to distinguish" [Millikan, 2013, p138]. So far, so good.

However, maxMI features another version of the reference class problem. Not only must we be able to limit the values *within* an eRV, we must be able to compare *between* eRVs. A given neural representation will typically have a very wide receptive field, especially so the further downstream the representation is. The wide receptive field means that the upstream discriminatory capacities of the organism fail to delimit just one eRV. Many possible eRVs, albeit each with outcome values limited by the discriminatory capacities of sensory systems, will be candidates for modelling the representation's content.

If a representation is limited to certain sensory inputs, this does not yet tell you the arrangement of those inputs, or the frequency of a given input, or whether the representation is limited to some subset of the inputs it has, in principle, in its receptive field. The collections of all the possible outcomes the representation *could* have in its receptive field is still potentially infinitely large. These define the space of all possible eRVs which could model the content of the neural representation.

The theory maxMI tells us that the relevant eRV is that eRV with which the representation, modelled as an iRV, has maximal mutual information. However, we need to compare values of mutual information between the potentially infinite set of eRVs. Which eRV the iRV has maximal mutual information with will vary depending on the reference class we choose and unless we choose a reference class, we can make no sense of comparing mutual information values with a potentially infinite set. We need to choose a reference class non-arbitrarily and non-pragmatically in order to find a scientifically acceptable representational content.

In this chapter, I argue that taking the 'C-function' of the subsystem to which the target representation belongs provides the relevant reference class of eRVs for comparing levels of mutual information. In short, we must limit the range of possible eRVs to those which could, in principle, meet the downstream information requirements for the fulfilment of a given cognitive function. Specifically, they must be those eRVs which could be *processed* by the system, which is served by the function of the system as understood according to Cummins' analysis of functions (hence C-function). I summarise the argument in section 6.5.

6.3 Functions

To characterise the information which initial systems provide to downstream systems, we can make use of an existing theoretical framework, that of **functions**. Functions provide a way of associating some subsystem to some external part of the world by way of identifying the role that the subsystem plays in enabling some capacity, of a wider system, which has to do with the world (e.g. recognising faces). We will be particularly interested in Neander's sense of information-carrying functions (see chapter two 2.4).

Information-carrying functions tell us which environmental item the system must be related to such that it is able to provide downstream areas with the information they need to fulfil some capacity. Without a function to carry information, a system may co-vary, however strongly, with some stimulus, but that co-variation will never be *used* by the system itself. If the system itself does not use the information, we fail to have any explanatorily relevant content. Functions bridge the gap between co-variation and content

proper. As argued in chapter three 3.3.2, we need functions in order to substantiate the idea that information is *encoded* within a system; only if there is something downstream which *decodes* the information is that information part of an encoding-decoding codex.

6.3.1 C-functions and W-functions

In the philosophical literature, we typically find two notions of function: an etiological notion, primarily defended by Wright - I will call these **W-functions**; and a non-etiological notion, primarily defended by Cummins - I will call these **C-functions**.

An etiological explanation is an explanation of the presence of something in terms of its history. This typically involves invoking either evolutionary or learning processes. For Wright, to ascribe a function to some part of a system is to explain the presence of that part of the system with respect to past adaptive contributions of that part to the wider system. Wright's formulation of W-functions is:

> The function of X is Z *means*
>
> (a) X is there because it does Z
>
> (b) Z is a consequence (or result) of X's being there. [Wright, 1973, p161]

Where (a) expresses the etiological character of function ascription. For example, we might elaborate on X being there "because" it does Z by way of providing an evolutionary explanation in terms of the adaptiveness of Z for a species, and, (b), the fact that members of the species having feature X enabled them to perform Z. It is adaptive for pigeons to fly (to escape predators, to scout for food, etc.), and having wings allows pigeons to fly. So, the W-function of wings is to enable flight (we might say: wings are

flight-enablers).

Cummins' non-etiological account prioritises Wright's condition (b) and ignores condition (a). As Cummins understands the term function, functional ascription serves to answer the question: *how* does the system perform Z? Cummins writes of functional analysis: "a what-is-it-for question is construed as a question about the contribution 'it' [the state or structure in question] makes to the capacities of some containing system" [Cummins, , p164]. Cummins also characterises this as a "how-does-it-work question" [Cummins, , p165]. Here is the definition:

> x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s's capacity to ψ just in case x is capable of ϕ-ing in s and A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to ϕ in s [Cummins, 1975, p762]

For some pigeon, take x as *wings*, ϕ as *flight enabler*, ψ as *fly*, s as *pigeon*, and A as an *explanation of pigeon flying*. The function of wings in pigeons is to enable flight relative to an explanation of the pigeon's capacity to fly, just in case wings are capable of enabling flight in pigeons, and the explanation of pigeon flying appropriately and adequately accounts for the pigeon's capacity to fly by, in part, appealing to the capacity of wings to enable flight in the pigeon.

This definition looks, *prima facie*, like a dormitive virtue account of flying. However, a fully developed explanation of *how* wings act as flight enablers overcomes this issue. The main point is that the C-function of some part of a wider system is the role that part plays in realising the capacity of the system as a whole. Cummins notes that, given this definition:

> If, for some reason, flying ceased to contribute to the capacity of pigeons to maintain their species, or even undermined that capacity to some extent, we would still say that a function of the wings in pigeons is to enable them to fly. [Cummins, 1975, p755]

So, the account is non-etiological: we can ascribe a C-function without regard to the adaptive evolutionary or learning history behind its presence.

Below, I will argue that the cognitive functions which are attributed to brain regions in cognitive neuroscience are C-functions. In section 6.6.2 I will also argue that, in virtue of being capable of *dys*function, C-functions provide the basis for misrepresentation.

6.4 The argument for C-functions

In this section, I argue that C-functions are the relevant type of function for the implicit theory of content in cognitive neuroscience, given extreme cases of pluripotency.

Peter Godfrey-Smith argues [Godfrey-Smith, 1993] that W-functions and C-functions are not in direct conflict - they speak to different explanatory aims. Godfrey-Smith is therefore a function *pluralist*; he argues that some component of a system can have both a W-function and a C-function. Each type of function can be invoked relative to an appropriate explanation. As Godfrey-Smith writes: "components of the system [can] have both Wright functions and Cummins functions, and some of the Cummins functions [can be] opposed to the Wright functions." [Godfrey-Smith, 1993, p15]. C-functions are invoked in explanations of how a system operates, while W-functions are invoked in explanations of why a system functions that way rather than another. We covered this extensively

in chapter two (section 2.5), in which we made the distinction between proximate and ultimate explanations. In those terms, Godfrey-Smith argues that C-functions are appropriate for proximate explanations, and W-functions for ultimate explanations. In chapter two I endorsed, and argued for, this claim.

W-functions are predominantly accepted as the relevant functions for teleosemantics. Typically, the reason given is that W-functions, by contributing to explanations of *why* some state or structure is there, provide a way in which we can demarcate dysfunction[1]. If the state is active in a context in which it was not historically adaptive, the state is misrepresenting, since the state is *dys*functional. Were the state W-functioning, it would only be active in contexts in which it was historically adaptive. It is not as clear, the thought goes, that C-functions can define a content in such a way as to allow for misrepresentation. Some state will be C-functioning even in cases in which the realisation of the capacity afforded by the C-function is detrimental to the system. I will address this in section 6.6.2. Before that, I want to present an argument for why we, with all our constraints, and against the teleosemantic tradition, should take C-functions as the relevant type of function involved in content determination.

It is not a novel view that function ascriptions in cognitive neuroscience are based on C-functions: Godfrey-Smith notes that "functional claims in these fields often appear to make no reference to evolution or selection" and that in these fields "the attractive account of functions has always been that of Robert Cummins" [Godfrey-Smith, 1993,

[1]There is some terminological nuance between malfunction and dysfunction; malfunction often suggests complete breakdown of functioning whereas dysfunction often suggests continuing but impaired functioning relative to some standard. However, this difference tends to be glossed over. The term 'malfunction' is typically used in the philosophical literature, whereas 'dysfunction' is typically used in the scientific and medical literature. I will use dysfunction throughout with its conventional associations, and if 'malfunction' is meant I will indicate as much in the text.

p7]. Since we are using cognitive neuroscience as our explanatory framework, for reasons spelled out in section 5.5, we will be following the practice of cognitive neuroscience. So, Godfrey-Smith claims, we should follow them in using C-functions. Indeed we will.

However, we can motivate Godfrey-Smith's claim beyond noting, as he does, that there is often no reference to evolution or selection in neuroscience. As we know, (problematically) pragmatic factors may influence linguistic choices which do not reflect implicit theoretical commitments.

I will argue that neither informational teleosemanticists nor the cognitive scientists upon whose work we build can make use of W-functions. Accepting for now that functions play a role in isolating content, there are instances in which content plays an explanatory role, and in which scientists ascribe functions, but no W-function ascriptions can be made. C-functions, on the other hand, can be ascribed. Indeed, this is likely to be a widespread phenomenon. The reason for this is the 'pluripotency' of the cortex.

The cortex being pluripotent means that areas of the cortex are "capable of assuming a wide range of cognitive functions" [Bedny, 2017]. This claim means more than that one component of the cortex can simultaneously support multiple functions (although this is true): it is also the claim that one component of the cortex can, over time, acquire new functions. Note that this time span can be very short. Pluripotency is not limited to phylogeny - flexible acquisition of new functions for a given region can happen within the lifespan of an individual.

An example of pluripotency is given by Bedny [Bedny, 2017]. Bedny reports that in blind individuals, areas of the visual system appear to be activated during mathematical reasoning, spatial reasoning, and language-related processing. For example, Bedny

reports that "visual cortices of blind but not sighted individuals are active while solving auditory math equations and activity increases parametrically as the equations become more difficult" [Bedny, 2017, p641]. Similarly for language-related processing, response profiles usually seen in areas associated with language processing in sighted individuals (frontotemporal cortex) are found in the visual cortex of blind individuals. In particular, in blind individuals but not in sighted individuals (even those in blindfolded controls), "visual cortices are sensitive to subtle manipulations of grammatical structure. For two sentences with nearly identical meanings and words, the sentence that is more grammatically complex (i.e., has a syntactic movement dependency) produces larger responses" [Bedny, 2017, p641].

It should be immediately apparent that functional ascriptions in such cases cannot be made on the basis of what is adaptive in evolutionary terms. The frontotemporal cortex in blind individuals supports linguistic and mathematical functions, suggesting that areas can be flexibly recruited for numerous functions depending on occurent requirements. Generally, the cortex can support functions acquired during ontogeny, meaning that an evolutionary explanation is simply false. Nonetheless, functions are ascribed in these cases, despite no evolutionary selection. This immediately rules out W-function ascription based on evolutionary history. However, functional ascription in these cases is entirely consistent with C-function ascription.

However, W-functions are not limited to evolutionary history. Wright's definition stipulates that W-functions can be acquired through *learning* history. If some function was produced by learning to enable the performance of y, and y is beneficial for the system as a whole, the W-function of the component is to do y.

It is not always clear which notion of learning is supposed to be relevant for the attribution of W-functions. However, I argue that we should use the model of learning employed by Shea - a form of reinforcement learning. This is because Shea uses it precisely as an example of how he considers learning in relation to W-function ascription, and Shea's account is arguably the most well-elaborated and well-motivated account of W-function ascription in the contemporary literature.

For Shea, a W-function can be ascribed to a process within a system if that process has been "stabilized" [Shea, 2018, p56]. For a process to be stabilised is for it to be reproduced in virtue of the fact that the process led to a beneficial outcome in the past. Shea pinpoints learning with feedback as one relatively low-level mechanism responsible for stabilisation [Shea, 2018, p59]. The most basic variety of learning involved in learning with feedback is reinforcement learning. This learning can be achieved with either positive or negative reinforcement [Shea, 2018, p62]. Some behaviour which produces a reward, where failure to repeat that behaviour produces a punishment, increases the probability that the behaviour will be repeated. This is the most basic sense of learning relevant for the ascription of W-functions (specifically, as functions to produce that behaviour).

The type of pluripotency described in blind individuals by Bedny could be considered the result of reinforcement learning. Bedny describes the mechanism responsible for this form of pluripotency as follows: "in the sighted, MT [a region of the visual system] receives low-level sensory input from primary visual cortices, whereas by hypothesis MT of blind individuals receives highly processed motion-related information from parietal cortices" [Bedny, 2017, p643]. It is this spatial information which is thought to enable

to recruitment of the visual system for mathematical reasoning and language processing. Potentially, feedback in the form of successful navigation, for example, may have reinforced the connection of spatial information to MT.

However, this is not a given. Bedny suggests a model according to which, during development, cortical areas are constantly fought over for information processing from numerous inputs. Bedny explains the idea with the following metaphor: functional specialisation is "a self-organizing process, where different inputs of information compete for cortical real estate" [Bedny, 2017, p645]. In this highly competitive environment, it is simply deprivation of one input which leads to the take-over of the region by another input. Nothing is reinforced: an obstacle is removed and an alternative input intrudes. So, in such cases, MT has the function to process spatial information, but has not acquired this function through learning in Shea's sense.

Another highly influential and much-cited model of pluripotency, due to M-Marsel Mesulam [Mesulam, 1990], suggests that cortical functions can change rapidly, with information channels directed to multiple higher regions on the fly. According to Mesulam, each low-level region, such as the regions we have been considering in the visual system, can be recruited quickly for myriad and veriegated downstream applications. Numerous feedback and feedforward connections found between regions support this process. Mesulam suggests that "retrieval can be initiated from any point" [Mesulam, 1990, p607] of the processing hierarchy we have been considering: connections to the frontal lobe throughout the system allow for contributions to be made to downstream areas at "all levels of complex processing" [Mesulam, 1990, p608]. Mesulam suggests that this process can be "spontaneous" [Mesulam, 1990, p597], requiring no reinforcement. To be

fully transparent, Mesulam says "spontaneous learning": however, whatever model of learning Mesulam has in mind, it is not one which fits with W-functions: there has been no previous instance of the connection with a specifiable benefit in virtue of which the connection is strengthened (it may never be repeated, in fact). The connection is made rapidly to serve an occurrent need, and there may be no stabilisation at all - the recruitment of the lower area may never be repeated. We cannot ascribe a W-function, but we can ascribe changing and constantly updating C-functions.[2]

Although one may suggest that the connection is made in *anticipation* of a future reward, Shea specifically rules out future-directed benefit in W-function learning. He argues that this over-generates, since "it is a very open-ended matter whether an output would contribute to the persistence of an organism, or would be stabilized by feedback-based learning, or would promote reproductive fitness" [Shea, 2018, p63]. Too many things *could* have some future benefit. On such a view, W-functions would be ascribed to near enough anything. Additionally, Shea notes that future-oriented W-function ascription undermines the role that W-functions have in causal explanations, since something that has not yet happened cannot causally explain why something has happened now [Shea, 2018, p63].

I maintain that the things components occurently contribute to the system as a whole, its C-functions, are what matters when it comes to cognitive neuroscience. If one were to plug sensory inputs into a part of the cortex, and that part of the cortex afforded the same capacities as the part of the cortex that the input was previously plugged into, cognitive neuroscientists would ascribe to that new part of the cortex the corresponding

[2]One curious result of this is that, very possibly, the content of both the higher- and lower-level representations may be in near-constant flux, depending on the relative levels of stability of the C-functions.

C-functions of the previous part of the cortex. In fact, this is very near to actual empirical cases. For example, in von Melchner et al. [Von Melchner et al., 2000], axons from the ganglion cells of ferret eyes were induced to connect to the ferret's auditory system. The ferrets were able to perform visual tasks requiring visual responsivity, with activity registered in the auditory system. von Melchner et al. therefore ascribe visual functions to the ferret's auditory system. This case is complicated by the fact that the connections had to be made during ontogeny for practical reasons: nonetheless, the specific learning history is entirely irrelevant to the study: what matters is what is connected to what.

This latter point is made emphatically by Hagoort and Indefrey, who write that"the basic principle of brain organization for higher cognitive functions proposes that these functions are based on the interaction between numerous neuronal circuits and brain regions that support the various contributing functional components" [Hagoort and Indefrey, 2014, p359]. According to Hagoort and Indefrey, there is simply no need to consider what happened in the past in order to ascribe functions. What matters is what cortical regions *do now*.

To conclude the argument: W-functions cannot be the relevant functions for our brand of informational teleosemantics. A systematic application of information theory requires that we attend to the functions of cortical brain regions to specify external-side random variables. The functions of cortical brain regions which have explanatory purchase are C-functions: W-functions have no explanatory purchase, since the downstream capacities thought to be enabled by functions can be enabled when there are no W-function ascriptions and only C-function ascriptions.

6.5 C-functions and initial eRVs

Given that C-functions are the relevant type of function for characterising the role a cortical region has in enabling a cognitive capacity, in this section I spell out exactly how C-function ascriptions delineate the range of relevant eRVs for comparison of mutual information values between candidate contents.

C-function ascriptions are made on the basis of how a component features in explanations of downstream capacities. One may explain a downstream capacity in a number of ways, at varying levels of grain, with various ways of describing the environmental entity which contributes to the capacity. This means that it is indeterminate precisely which phenomenon is picked out by such explanations. As in life more generally, there are always a number of ways to achieve the same result.

There are many aspects of the environment which might be sampled by the system to enable a capacity. To take a very simplified example, there are levels of visual acuity which are all consistent with enabling the capacity of picking up a mug. I can pick up a mug with my glasses on or off. Given a high level of visual acuity perhaps I am able to represent $\langle$*mug handle*$\rangle$. With a low level of visual acuity I may only be able to represent $\langle$*vaguely handle-like-thing*$\rangle$. Depending on which representation I recruit, my capacity is better or worse. However, if the mug is close enough, I can pick it up using either representation. In this example, two different contents have enabled the same cognitive capacity.

Nonetheless, explanations of my capacity to lift mugs *constrain* the range of external entities which are relevant. This provides us with a way of limiting the range of possible external entities which can serve as content. A range of possible external random vari-

ables is established, with each random variable modelling an entity which is a candidate for the content of a target neural structure. This range of eRVs models all those entities which are *potentially* explanatorily relevant to the system itself, where indeterminacy results from the fact that multiple items could, theoretically, support the same capacity. Of course, we then need some further constraint to isolate the item which is *actually* used by the system. This is precisely what the relation of maximal mutual information offers - the content of the representation - the item actually used to perform a given cognitive capacity - is that item modelled by the eRV which has maximal mutual information with the iRV used to model the representation.

The parameters of the eRV will be determined, as we saw in the previous chapter, by the parameters of the sensory systems which have the information-carrying functions we are ascribing and the probable distribution of outcomes in the environment. Functions allow us to limit the potentially infinite range of possible eRVs to just those potentially explanatorily relevant for the capacity under investigation - those which are content-candidates.

C-functions - as opposed to W-functions - serve this role particularly well. They require us to isolate *usable* information (see section 7.5.3). I claim that, in order for an account which uses representation to "appropriately and adequately account" for a capacity (as described in the definition of C-function - section 6.3.1), where this involves the fact that the representation is capable of *doing something* (ϕ-ing) within the system, the representation ought to have some identifiable *means* of doing that thing. There needs to be some way in which the representation interacts with the downstream structures which enact the capacity.

With the focus firmly on what the component *can* occurrently *do* within the system, C-functions require us to find which information values are *actually* processed by the system itself. This is crucial for maxMI, since it requires that content contain information *available* to the system itself to be explanatorily relevant. The next chapter is an investigation of maxMI, and how it secures availability, relying heavily on C-functions to do so.

6.6 Objections to C-functions

In this section I consider why C-functions have been thought inadequate for an account of content. Two concerns are particularly important for us to address. The first is that C-functions a problematically pragmatic - they are based on the arbitrary aims and interests of researchers. Given our generally naturalistic and realist project, this would be a problem, as it would delimit a range of eRVs based on problematically pragmatic features. Thus, our content ascription would be problematically pragmatic.

Second, C-functions are thought to rule out the possibility of dysfunction. This is a problem due to the fact that we are attempting to find an implicit theory of content in cognitive neuroscience, while cognitive neuroscience - including those regions which meet the criteria set out in chapter three (section 3.5) - implicitly invokes dysfunction. A common experimental paradigm involves generalising from dysfunction in lesion patients to functioning in healthy individuals.

So, we need to answer each criticism to be sure that C-functions can serve the role we have argued they have. I take each in turn.

6.6.1 Pragmatism revisited

Neander argues that, since the definition of C-function directly appeals to "a researcher's explanatory aims" [Neander, 2017b, p54], C-function ascription is sensitive to a "pragmatic determination of the boundaries of the system" [Neander, 2017b, p55]. How we break up the system into parts is thought to be pragmatically determined. This would contravene the no-pragmatism requirement we are working with. The basic problem is that the content of a given representation would change depending on what we, as observers, choose to use it to explain. This would be a change in content which does not result in a change for the system itself. *This* is the main concern facing us with respect to pragmatism.

Thankfully, we can appeal to the distinction between problematic and non-problematic pragmatism introduced in the previous chapter (section 5.6). Non-problematic pragmatism amounts to developing a hypothesis about which elements of the system are relevant to the capacity under investigation. So, if our interest is in explaining flight, we will select wings for analysis, under the hypothesis that they are relevant for flight. In this case, selecting wings as our unit of analysis is not pragmatic except in the sense that isolating any phenomenon for the purposes of any explanation is pragmatic. Selection is not guided arbitrarily, but by a hypothesis about what is relevant for a given capacity.

However, a further worry is that there is a choice to be made between what we happen to be interested in explaining - which capacity we choose to study. If we want to explain the contribution of wings to flying, we would have to ascribe a different C-function to wings from the C-function we would have to ascribe if we want to explain the contribution of wings to egg warming, for example.

Ori Hacohen makes this very point. Hacohen supports the view, outlined in section 6.4, that C-functions are the relevant functions for cognitive neuroscience, but bites the bullet; representations as used in neuroscientific explanations are constitutively dependent on the mental states of the scientists performing those experiments. Hacohen writes:

> CFA [Cummins Functions Approach] violates the naturalistic constraint by defining representations relative to a given explanation and with respect to scientists' explanatory aims. [Hacohen, 2022, p715]

What this means, specifically, is spelled out in a quote from Neander which Hacohen endorses:

> There are explanatory aims when anyone tries to explain complex or, for that matter, simple capacities. And which causal contributions ought to be mentioned in a given explanatory context will depend on one's aims. But, on Cummins' account, if there are no relevant explanatory aims, then there are no functions. Explanatory aims are constitutive for Cummins functions. [Neander, 2017a, p710]

Neander is suggesting that there are no C-functions without actual, existing explanations. And if there are no C-functions without explanations, and content is determined in part by C-functions, there is no content without an explanation. But if explanations are driven by pragmatic conerns (such as what we are merely interested in), we have a pragmatic account of functions, plunging us back into Egan's concern that content is

determined pragmatically. Content would therefore fail to be part of the theory proper of cognitive science, contrary to what we have been trying to establish.

So, how should we respond to Hacohen and Neander's charge? The obvious route is to point out that explanations, insofar as they are accurate, pick out genuine phenomena which exist independently of those explanations. When Cummins suggests that "the function of x in s is to ϕ relative to an analytical account A" we need not read this as suggesting that A must *actually* have been committed to paper, or have been dreamt up by any particular scientist. The analytical account in question is an explanation of a capacity. If x does ϕ in s this is so independently of whether we choose to explain it: the relevance of citing the explanation lies in pointing out that x *actually* 'does something' in s, and that the explanation A cites this phenomenon.

This is why Cummins notes that A must "appropriately and adequately account" for a capacity ψ of s by citing x's capacity to ϕ. While a full story about what makes an 'appropriate' or 'adequate' explanation is elusive - and we lack the space to consider it - what should be clear is that Cummins' qualification should be understood as requiring that the explanation has the same force as any other scientific explanation, and such explanations are not intended to rely on the very explanation being given for the explanation of the phenomenon they are trying to explain. Or, if they are, non-naturalism is rampant in science and not a unique concern for content theorists.

Hacohen is alert to this line of response. However, while he agrees that the explanations in question *do* specify actual phenomena, he argues that the phenomena themselves are not specifiable in terms of their *representational* contributions to higher capacities except in relation to an explanation. As he puts it:

> In other words, the neuroscientific phenomenon, while objectively real, isn't objectively given. Nothing objectively singles out this particular phenomenon from others, at least not to the extent that is necessary to define a determinate neural representation. [Hacohen, 2022, p711]

The idea is that, depending on which capacity we *choose* to explain, the very same component will have different C-functions, and so different content attributions. Hacohen elaborates on the point by way of raising two indeterminacy challenges. The first involves suggesting that the same structure can contribute to two or more very different capacities. This would mean that the very same system could have multiple C-functions.

This certainly is a problem if we are committed to the idea that a component can only have one C-function. I see no reason to assume this principle. Provided we ascribe C-functions non-problematically pragmatically, one component can *surely* have multiple *real* C-functions. It is in principle possible that we could pursue the explanatory question: how do pigeons manage to keep eggs in the nest warm? It is in principle possible that lowering the wings in the nest contributes to warming the eggs. So pigeon wings might have two C-functions: to enable flight, and to keep eggs warm. Which one we choose to *focus on* will depend on our explanatory aims, but which C-function the wings *actually have* is just dependent on their role in enabling either capacity, independently of whether we notice the function or not.

Hacohen's further point is that if we attribute C-functions "*relative to a specific phenomenon*, we define it in a manner that is dependent on our explanatory aims" [Hacohen, 2022, p712], we render *content* (not just functional ascription) dependent on our explanatory aims. However, we can respond in the same way as above: we allow that a single repre-

sentation can have multiple contents, where those contents are real aspects of the system, and not dependent on our interests - we only *uncover* the contents when we investigate the role the representation plays in enabling multiple capacities.

Take Hacohen's helpful example of the orientation-selective cell: "the same neurons in V1 that enable orientation detection also enable contrast discrimination, and are simultaneously sensitive to both orientation and contrast" [Hacohen, 2022, p711n]. If this is true, then it is possible that these cells have at least two contents, something like $\langle line \rangle$ (with certain orientations, as we have discussed previously), and $\langle light\ contrast \rangle$[3]. But, *pace* Hacohen, these are not contextually defined by what we choose to explain: they are both contents which can be accessed *by the system itself*[4]. The relevant context for fixing content is not our explanatory aim, but which content is being accessed by which downstream systems. Probably, these states are indeterminate with respect to each content, with different downstream systems consuming different contents from the same state at the same time.

Scientists may talk about 'the' content of the state when we advance a theory which focuses on just one downstream capacity. This is loose talk - we have no reason to attribute it to part of the theory proper. What they presumably mean is something like "the content which is relevant to our current explanatory aims". This does not mean that the state fails, in general, to specify any other content. This is a standard case of highlighting just one aspect of a complex system for the purposes of explanation, while abstracting other aspects. This is a perfectly standard scientific practice, not unique to

[3]Of course, if we were seriously positing contents we would be doing empirical work and using operationalised terminology. These are just toy examples.

[4]See chapter seven section 7.5 for more on availability of information.

cognitive neuroscience. It has no bearing on the reality of either the attended to nor the neglected aspects of the system.

Hacohen's second charge of indeterminacy is that, even relative to one capacity, the content is indeterminate between the various aspects along the causal chain. Hacohen suggests that we invoke explanations to determine which part of the causal chain is relevant, since which aspect of the causal chain we are interested in is picked out by the explanation. However, Hacohen argues, this renders content relative to our explanatory aims. The content the state has is fixed by whatever part of the causal chain we take to be relevant in our explanation.

There are two responses to be made: first, the content may genuinely be indeterminate between aspects of the causal chain if the explanation of the cognitive capacity in question could be made given only proximal regions. However, the explanations we are looking at in this thesis, *do* call on just the distal part of the explanatory chain since, by design, they involve external items as part of the explanation (e.g. face recognition). But, I argue, if that explanation is accurate, the system itself will *actually be using* information about just one part of that chain. Take as an intuitive example the following: we understand the sense in which there are limitations on what content from lower levels is available to higher levels[5] just by reflection on our own higher-level representations; I cannot now call to mind the precise angles of the edges of my keyboard in front of me relative to my retina. However, some part of the causal chain which gets that information to me 'has' that information. I, whatever I am, cannot access it - not directly, at least.

More formally, citing some content in an explanation of a capacity must commit one

[5]Provided that the hypothesis that higher-level representations depend on lower-level representations for their content

to the fact that the information about more proximal states is not, as matter of fact (not as a matter of interest), used by the system. We must be committed to the fact that information specifying a distal element is used. According to maxMI, the distal element used by the system is within the range of eRVs delimited by the C-function of the subsystem containing the representation.

Spurious functions

Another worry is that we may frivolously ascribe spurious C-functions. For example, based on pigeons' capacity to make a whooshing sound as they fly, we may ascribe the C-function of 'making a whooshing sound' to wings. We can make two possible responses. First, we allow such C-functions (alongside the other C-functions), and chalk it up to various components of the universe having some *potential* to do *something*. Such C-functions are currently entirely uninteresting, but who knows, perhaps some change in the pigeon's environment in the future will mean that making a whooshing sound becomes very important - perhaps some new predator is scared of that whooshing sound and the C-function comes in handy.

Alternatively, the response I favour, we could invoke Cummins' notion of the in-order-to relation: C-functions have their place in a hierarchy of embedded capacities. Everything we are typically interested in has some terminus in serving some primitive requirement of the organism: each capacity serves wants, needs (hopes, dreams). If we can find a theoretically principled way to find a terminus for the in-order-to-relation, we might be able to limit the ascription of C-functions to non-frivolous capacities on the basis that things like whooshing have no place in the hierarchy of embedded capacities.

On this account, whooshing could *become* a C-function of wings if it suddenly warded off predators, but only once it *actually* did so.

Invoking the in-order-to-relation has the benefit that we isolate just those C-functions which make a difference for the system itself. If whooshing serves no need of the pigeon, then if wings were to cease to make a whooshing sound, it would make no difference for the pigeon. However, since flying serves several needs of the pigeon, it makes a big difference if wings cease to be flight-enablers. The C-function of enabling flight is a function the possession or absence of which makes a difference for the organism itself, while the C-function of whooshing is not. Given that we are looking for content ascriptions which make a difference to the system itself, we are motivated to limit C-function attributions to those which serve some need of the possessor of the cognitive system.

However, I do not see any particular difficulty in accommodating either tactic, on the assumption that components *can* have multiple C-functions, even those which are explanatory for frivolous capacities. The point would be that only some C-functions are involved in content determination within the system, so no unacceptable proliferation of content would result.

6.6.2 Is there C-dysfunction?

In section 6.4, I posed the question whether components with C-functions can be dysfunctional or malfunction. Malfunction is understood to be a complete failure to function and dysfunction is understood to be partial, impaired[6] functioning. In this section I will

[6]For C-functions, we should understand 'impaired' relative to enabling the capacity in the context of which we attribute the function. In principle, an impaired C-function may enable a different capacity and may even generally 'improve' the system as a whole with respect to some other set of capacities. A bad flight-enabler may make a fantastic egg warmer.

answer in the affirmative.

Why are dysfunction and malfunction important? One answer, typically given in the literature, is that dysfunction grounds error. We know all too well that we are often wrong. We have many false beliefs which generate false expectations, cause us to utter false sentences, and generally make us look foolish. We would like to have a naturalistic account of this ability to be wrong. Being wrong is not a given: it is an achievement. Rocks are never wrong. Arguably, basic organisms and plants are never wrong. Some plants *might* be said to be wrong: a fly-catcher might be wrong when it catches a falling leaf rather than a fly, but without a theory of dysfunction we cannot say for certain whether it is wrong or just doing something *we* know not to be in its best interests. Further, since the gift of inaccuracy is often taken to be the mark of a non-naturalistic component of representation, we will only complete the naturalistic project once we have given it a proper account. Malfunction and dysfunction are often thought to fulfill this role; they are ways, as Dretske puts it, for nature to be wrong.

However, on the face of it, if some component does not allow for some capacity to be realised, that component simply has no relevant C-function. However, dysfunction and malfunction require that the component *have* the relevant C-function, but fail to perform its role in realising the capacity (or do so in an impaired way). As an intuitive example: it makes very little sense to speak about my eyeball malfunctioning because it does not allow me to fly. It is true that my eyeball does not do that, but my eyeball never had the function to do that in the first place. So, it looks like C-function cannot ground error because they cannot be dysfunctional.

Whatever our intuitions, it is important to know whether C-functions can dysfunc-

tion, since error can play a vital role in scientific explanations. A widely used paradigm in neuroscience involves taking a patient with some cognitive dysfunction and making inferences about non-dysfunctional individuals. For example, Almeida et al. [Almeida et al., 2020] investigate the nature of face-space representations by conducting a study involving a patient with Hemi-prosopometamorphopsia (hemi-PMO), a condition in which "brain-damaged patients perceive one side of the face as distorted, with features that appear out of proportion, drooping, or swollen" [Almeida et al., 2020, p4071].

Almeida et al. use hemi-PMO to discover whether the "reference frame" of the face-space *in healthy patients* is "retino-centered", "stimulus-centered" or "face-centered".

Retino-centered distortions appear on the face depending on where on the retina the face is projected.

Stimulus-centered distortions always appear on the same side of the stimulus, regardless of where on the retina it is projected *and* regardless of the orientation of the stimulus (e.g. if the stimulus is upside-down, the distortion appears on the same side of the stimulus relative to the viewer).

Face-centered distortions, however, always appear on the same side of the stimulus *except* if the stimulus is oriented in a different direction, in which case the distortion moves with the stimulus (e.g. an upside-down face originally distorted on the left relative to the viewer is now distorted on the right relative to the viewer).

Almeida et al. found that hemi-PMO appears to be face-centered; the distortions track the orientation of the face.

Since it is hypothesised that "Hemi-PMO results from disruptions to representations coded within a particular reference frame" [Almeida et al., 2020, p4071], Almeida et al.

conclude that "the human visual system contains procedures that encode faces in a face-centered frame of reference" [Almeida et al., 2020, p4073]. This means that "representations of different in-depth rotations are aligned to a common template" [Almeida et al., 2020, p4073] in those suffering from hemi-PMO and those not alike.

The important point for current purposes is that hemi-PMO is understood as a *disruption* of ordinary functioning. Almeida et al. hypothesise that hemi-PMO results from "a disruption to information transmission from one hemisphere to the other" [Almeida et al., 2020, p4074], suggesting that the information transmission in hemi-PMO patients does not function as it does in healthy individuals. The explanatory value of dysfunction in this case rests in the fact that we can make inferences about non-dysfunctional information transmission. Given that hemi-PMO results from dysfunction, we are in a position to know what proper functioning involves, since there *is* a function that the information-transmission serves - it is simply not serving its function in patients with hemi-PMO. If hemi-PMO did not involve dysfunction, it would not be informative with respect to non-hemi-PMO cases.

Thankfully, despite appearances, C-function attribution can be made sensibly in cases in which the component with the C-function is currently failing to realise the capacity in virtue of which we attribute the C-function. Godfrey-Smith points out [Godfrey-Smith, 1993, p7] that this requires the invocation of a type/token distinction. C-functions pick out the role that the *type* of component has within the *type* of system we have before us. This *token* component in this *token* system is not fulfilling its C-function, but in virtue of having that C-function in other systems of the same type, it still has the C-function. So, it is dysfunctional or malfunctioning.

As Ori Hacohen highlights, this has raised the worry that "Cummins functions will run into trouble once we try to describe how the function of a type is determined" [Hacohen, 2022, p15]. For instance, as an example of how we attribute C-functions to a *type* rather than a *token*, Justin Garson [Garson, 2019] suggests that the statistical prevalence of the causal role of a component within other systems provides the relevant type. If, in a statistically significant number of cases, hearts pump blood, then hearts have the C-function of pumping blood. A token heart failing to pump blood still has the C-function of pumping blood, because most other hearts pump blood.

Garson goes on to dismiss the idea, since "if everyone's heart seized up at once, nobody's heart would have a function anymore, so nobody's heart would be dysfunctional" [Garson, 2019, p7].

Hacohen replies that Garson misses the fact that C-functions rely on the explanatory role that components play in virtue of which they have functions attributed. He notes that such explanations are "not necessarily dependent on the percentage of tokens that exhibit this effect" [Hacohen, 2022, p15]. Hacohen's suggestion is that we individuate types with respect to the explanatory role that this *type of component* plays in other systems to *enable the same capacity*. For example, information transmission between hemispheres, enabled by "the fibers that traverse the splenium" [Almeida et al., 2020, p4074], in those without hemi-PMO, enable mapping to a face-centered frame of reference, as they do in patients with hemi-PMO - albeit in a dysfunctional way.

However, Garson's counterexample involves *no* hearts pumping blood. Hacohen should require that at least *one* heart features in an explanation of pumping blood if we are typing hearts with respect to the role the heart plays in that explanation; if the

heart plays no role in *any* explanation of blood-pumping, there simply is no explanation involving that type of component for any other system, so we fail to identify the relevant type.

Instead, I argue that we should invoke a counterfactual. [7] We should maintain that, given an analysis of a system, the component *would* play the relevant role *were* it so embedded in a system which *did* have the relevant capacity. Imagine being an alien scientist visiting Earth following a global heart seizure. The alien scientist would be able to work out the C-function of the heart by running various analyses: electrocuting the heart *does* cause blood to pump around the system, so the heart sufficiently supplied with energy *would* pump blood around the system. So, this must be the C-function of the heart.

This is not an idle thought experiment. It is how neurophysiologists actually work out the C-function of a component. In order to discover the C-function of ganglion cells, scientists construct experimental conditions which involve "subjecting the retina to known intensities of light," following which "an electrical response can be elicited and recorded" [Mead and Tomarev, 2016, p10]. In fact, most methods involve this basic try-it-and-see approach (see [Mead and Tomarev, 2016] for a review); if the organism *were* in such-and-such a situation, what *would* this component do?

This method involves identifying a component independently of its function. However, Bence Nanay argues that there is circularity inherent in identifying a type of component for non-etiological theories of function [Nanay, 2014, p803]. We cannot type-identify the component in terms of its function. For example, if we type-identify hearts

[7] Neander makes this suggestion on behalf of the C-function advocate [Neander, 2017a, p11].

as blood-pumpers and this thing before us is not currently pumping blood, then it is not a heart, since - so the argument goes - it does not have the non-etiological function of pumping blood, since it cannot do it. However, if we say that it has the function but is not currently performing it, then we need some independent account of what makes this the same type of thing.

Here we can appeal to neurophysiological practice, which involves finding the **phenotypic structure** of a component and type-identifying it with respect to that structure. The same physical structure, resulting from the genetic code, across components, groups those components into the same type. For example, type-identifying ganglion cells involves a battery of tests which assess sameness of phenotypic structure. As Mead and Tomarev write, they are looking for "good phenotypic markers" which "leave no doubt to the identity of the cell in question" [Mead and Tomarev, 2016, p2]. The tests involved are highly technical, but essentially involve finding which compounds stain the target cells, and how much those compounds stain those cells. These factors tell researchers whether the cell is a ganglion cell based on the known absorption profile of the ganglion cell (due to its phenotypic structure). So, neurophysiologists can independently assess which type of cell they are dealing with, *then* find how that cell responds under various conditions in order to attribute a C-function to that cell type.

We can overcome the circularity Nanay points to if we individuate cells with respect to their phenotypic structure, and attribute a C-function on the basis of what that cell does when embedded in a system of a given type, in response to various experimental conditions. The system itself is to be identified, as Hacohen suggests, in virtue of meeting the same explanatory criteria as some target system, such as a system performing the

same cognitive capacity with the same hardware.

6.6.3 Section summary

To summarise, we can re-interpret Cummins' original definition:

> x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s's capacity to ψ, just in case x's are capable of ϕ-ing in s and A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to ϕ in s

Such that in "an analytical account A of s's capacity to ψ" s refers to relevant systems, either the specific system under investigation, or those in which something sharing a phenotype with x contributes to the type of capacity which is ϕ-ing. The explanatory value of this amendment is that it allows inferences to be made from dysfunctional cases to non-dysfunctional cases.

6.7 Swamp-people evaded

A benefit of using C-functions is that our brand of informational teleosemantics completely evades classic Swamp-person counterexamples. If a being, atomically identical in every way to an existing human, were created immediately by a flash of lightning in a swamp, it would have all the same C-functions as the existing human; all of its neural circuitry would be the same and its components would share the same phenotypic structure[8]. As Nanay writes:

[8]If swamp-people have no genes, they are not atomically identical. That is, unless we identify genes according to their history, which would not be in the spirit of a non-etiological account to begin with. I

> the swampman problem (if it is indeed a problem) is a direct consequence of the etiological account of function that teleosemantics tends to use. If we can use a different account of function in teleosemantics, this problem (again, if it is indeed a problem) would just go away [Nanay, 2014, p801]

Nanay hedges, but swampman *is* a problem for any account which attempts to answer how-questions. If we want to explain a cognitive capacity in terms of content, and swampman has the very same cognitive capacity (e.g. swampman can recognise faces), then we *need* to ascribe content to swampman. Otherwise, content does not in fact do the explanatory work we thought it did: we could explain the same cognitive capacity, in physically identical systems, without it. Luckily, we evade the problem completely by using C-functions.

I also maintain that explanations aimed at answering why-questions *do not* face the swampman counter-example. As Millikan writes [Millikan, 2017], if we want to explain *why* swampman shows some behaviour, there is just no good answer. Swampman is a bit of a fluke himself, and does things for *no* historical reason. Swampman is only a problem for accounts of content aiming to provide answers to how-questions. But, I argue, those accounts should use C-functions, hence evade the problem.

6.8 Conclusion

Functions are required to narrow the range of possible eRVs in order to make sense of comparisons of amounts of mutual information. They also provide the basis for misrep-

understand, in this context, a gene as a physical structure within the DNA of the organism.

resentation.

In the previous chapter (section 5.5), we argued that we need to follow cognitive neuroscience in our search for an implicit theory of content. In this chapter, we argued that cognitive neuroscience uses C-functions; function ascriptions in cases of pluripotency are not committed to a model of learning which is consistent with W-function ascription. So, cognitive neruoscientists ascribe functions in cases in which W-function ascriptions cannot be made. Instead, they make functional ascriptions which are not only consistent with C-functions, but which mirror the ascription on the basis of contribution to downstream capacities. We therefore took C-functions as the relevant function for content-determination.

We then responded to the worry that C-function ascriptions render content constitutively dependent on the explanatory aims of scientists. We argued that scientific explanations pick out phenomena, and it is those phenomena which are picked out as C-functions, independently of whether any explanation is in the head of any scientist. We discussed Hacohen's response that this leads to indeterminacy.

In section 6.6.2, we considered whether components with C-functions can dysfunction or malfunction. We introduced Godfrey-Smith's point that C-function ascriptions are made using a type-token distinction. We then discussed the worry that there is no principled way of determining which other systems, or which counterfactual situations, are relevant to determining the type of the function. We followed Hacohen in arguing that the relevant systems are those in which the same component contributes to an explanation of the same capacity. We noted that there is a worry about how to type-identify the component itself. We saw that neurophysiological practice involves identifying com-

ponents in terms of their phenotypic structure.

In section 6.5, I argued that C-functions allow the delineation of a range of eRVs which are candidates for content. C-functions result in some indeterminacy: a host of possible items could *in principle* provide an explanation for some cognitive capacity. However, they are limited to those items which actually interact with the system itself. So, they isolate a range of possible eRVs which are potentially explanatorily relevant for content, allowing a meaningful comparison of mutual information values. We thereby avoid the reference class problem.

In the next chapter, armed with the constraints we have picked up along the way, I finally introduce maxMI and the main argument in its favour: it is presupposed by contemporary neuroscientific methodology (and is thought to be aimed at by classical neuroscientific practice).

Chapter 7

maxMI

7.1 Introduction

In previous chapters, we sought to discover the constraints on finding and providing the implicit theory of content in cognitive science.

In chapter three, we introduced three guiding principles to help narrow the search for an implicit theory of content to those regions of cognitive science in which content features in the theory proper (section 3.5). Broadly, these principles aim to discover studies in which content attributions pick out items changes in which result in changes for the system itself.

Then, in chapter four, we saw that contemporary studies, consistent with the three principles, use Shannon's information theory as their background theoretical framework (e.g. section 4.4). We saw how information theory provides a powerful tool for isolating the precise elements of the environment which are used by the system to perform

cognitive tasks.

We then paused, in chapter five, to consider how to carefully and systematically apply information theory to the cognitive system in light of Shannon's warning. Shannon points out that information theory, as a branch of mathematics, places certain constraints on the system to be modelled (section 5.2). Specifically, we saw that applying the random variable model requires isolating specifiable outcomes and calculating probabilities over those outcomes (summing to unity) (section 5.3). I argued that internal random variables, iRVs, can be specified by taking those outputs of a cell - in terms of its rate of firing, for example - which downstream regions are causally sensitive to (section 5.3.1). I argued that external random variables, eRVs, can be specified by taking those outputs of external items which the system's sensory interfaces are causally sensitive to (section 5.4).

However, I then argued that causal sensitivity to specific outcomes does not by itself isolate the relevant range of eRVs which model candidate contents of the target representation (section 6.2). There are potentially infinite possible ways to arrange the basic building blocks of things systems upstream of the target representation are causally responsive to - such as colours, lines, shapes, and so on - which each define a unique eRV. This set is potentially infinite. We need some way to limit the set of eRVs which are at least potentially content candidates - i.e. which could be used by the system for enacting a cognitive capacity. I argued that the C-function of the system containing the representation provides the requisite delimitation of the space of eRVs (section 6.5). Focusing on the role the representation actually plays in the system with respect to a given cognitive capacity limits the range of possible eRVs to those which can, in principle, serve the requisite function.

Now, in this final chapter, I set out maxMI. I argue that - according to the implicit theory in cognitive neuroscience - the content of a representation is the item in the environment which, modelled as an eRV, shares maximal mutual information with the representation, modelled as an iRV, relative to the other items in the eRV set delinated by the function of the system containing the representation.

I begin by considering existing information-based accounts which use "correlational information" - information the system has just in virtue of some representational state raising the probability of the presence of an item. I consider why theorists, such as Shea and Martinez, use correlational information rather than mutual information. I conclude that correlational information is a measure aimed at providing ultimate, rather than proximate, explanations (section 7.2.1). I also distinguish maxMI from a complimentary, but crucially distinct, theory, infomax (section 7.2.2).

I then provide an overview of mutual information as such, and spell out maxMI again 7.3.1.

Following this, I present the main argument in favour of maxMI being the implicit theory of content in cognitive science. I argue that the methodologies used in contemporary cognitive neuroscience to discover what a cell represents - namely, the spike-triggered average, maximally informative dimensions, and CMI - presuppose that the content of the representation is given by the item with which the representation has maximal mutual information (section 7.4). I also argue that the C-function of the system is used to initially limit the search range of candidate contents (section 7.4.4).

Finally, I set out how maxMI succeeds in providing content attributions for the system itself. We have placed this constraint on a theory of content throughout, and in section

7.5 I argue that maxMI isolates items about which the system has "available" information. Items about which the system has available information are those items outcome values of which receive processing by internal mechanisms (section 7.5.3). So, given the various constraints set out throughout the thesis, maxMI gives us a theory of content which can be used in the theory proper of cognitive science.

7.2 What measure of information?

Mutual information (MI) is the *prima facie* best tool to model the informational relationship between an iRV and eRV. Mutual information is a measure of the statistical dependence between two random variables, providing "an exact value for the strength of the association, measured in bits" [Stone, 2015, p134]. In intuitive terms, it tells us how much information about some entity is carried by a representation.

This looks incredibly helpful for working out the content of a representation; if a representational state contains a *huge* amount of information about milk, and a *really really tiny* amount of information about glass (say, because milk - very rarely nowadays - comes in glass bottles), that seems at least *relevant* for working out whether the representation has the content $\langle milk \rangle$ or $\langle glass \rangle$ - whatever one ends up saying.

So why is mutual information almost universally neglected by content theorists? Before I make the positive case *for* using mutual information (specifically, *maximal* mutual information) it will help to see the case in favour of a looser metric. Partly, because it means applying some familiar distinctions from chapter two (2.5) regarding differing explanatory projects. Also, from the outset I can distinguish maxMI from a similar-looking

theory called 'infomax', which maxMI is apt to be confused with. This should help set the ground for introducing maxMI proper.

7.2.1 Correlational information versus MI

In order to describe the informational relation between content and representation, Shea uses a measure he terms "correlational information" [Shea, 2020, p407], defined as:

$$p(Gb|Fa) \neq p(Gb) \tag{7.1}$$

In which G is some state of the environment, b, and F is a state of the cognitive system, a. The relation of correlational information thus holds between a representation and a state of the world provided the probability of the representation being present is not independent of the probability of the state of the world being present.

Neuroscientist Randy Gallistel suggests that Shea use mutual information instead. Gallistel argues that mutual information provides a "clearer, more intuitive, quantitative, and scientifically useful theory" [Gallistel, 2020, p3] than the model Shea uses. However, Shea maintains that he does "not want or need a stronger notion of correlation" than correlational information.

Shea's characterisation of correlational information is equivalent to Manolo Martinez's characterisation of "indication" [Martinez, 2013, p6], formalised as:

$$P(F|on) > P(F) \tag{7.2}$$

where F stands for the presence of some environmental item, and 'on' (versus 'off')

indicates the state of an internal mechanism (a representational vehicle, such as an activated neuron)[1].

So why do these authors, and others in the informational teleosemantic literature, use correlational information when a stronger measure of information is available (mutual information - see section 7.3)? To answer, we first need to go into a little detail on Shea's positive proposal:

> Contents are fixed relative to task functions. Task functions arise as a result of some stabilizing process. Learning is a key case. I argued that outcomes that are the target of stabilizing processes are often stabilized and robustly produced as a result of internal mechanisms, mechanisms that make use of exploitable relations between internal components and the world. [Shea, 2018, p217]

"Exploitable relations" refer to the correlational informational links between states of the cognitive system and the external environment. A task function is defined as some outcome that the system produces which has been "stabilised" by some process, such as evolution or learning [Shea, 2018, p65]. A mechanism with a given task function "exploits" a correlation between an internal state (to which the mechanism is connected) and the external environment.

Imagine some state is correlated (however loosely) with the presence of a predator. When the state is active, the organism engages in evasive behaviour. Thus, the correla-

[1]Superficially, Martinez's focus on the 'on' state appears to neglect the role inhibition can have for indicating some environmental item. Shea's formulation explicitly includes inhibition, since any dependence - in either direction - provides correlational information. This apparent difference does not run very deep: being 'on' indicates a 'hit', which can be spelled out however one likes in physiological terms, including activation and inhibition. This renders the two accounts equivalent.

tion has been exploited to enable the organism to, ultimately, stay alive. For Shea, the content of the internal state is the predator with which it is correlated, since this provides an "unmediated explanation" of the organism's behaviour [Shea, 2018, p84]: the least mediate explanation for how the outcome the correlation is exploited by was stabilised is that it was avoiding a predator[2]. We needn't mention any further correlations between predators and another item in the explanation.

Based on his positive proposal, there are two reasons why Shea rejects mutual information. First, such a notion may rule out the most immediate explanatory item. For example, an organism may have more mutual information with proximal properties. However, as Shea says, he aims to "explain how outputs were produced robustly and stabilized by interactions with the environment" [Shea, 2018, p90]. As covered in chapter two (section 2.5), this aligns with Mayr's conception of an ultimate explanation. Shea is addressing the question, roughly, why a given representation is present within the system.

In this sense, mutual information is not a helpful tool for answering such questions. It will produce content ascriptions which are not obviously related to adaptive benefit, except in a mediated way - we would need to offer additional environmental correlations to explain why an organism represents contents picked out by mutual information. However, given the explanatory role cognitive neuroscientists ascribe to content, i.e. aimed at answering proximal how-questions, this reason to reject mutual information is not relevant to the current project which aims to uncover the implicit theory of content in cognitive neuroscience.

[2]See also the discussion in chapter two section 2.4.1.

Second, Shea requires that there be an informational link which *can* be, but may not *currently* be, exploited. In other words, Shea's account emphasises what Cummins calls "unexploited content" [Cummins, 2010, p124], which is "information or content carried by or present in a representation that its harbouring system is, for one reason or another, unable to use or exploit" [Cummins, 2010, p122]. For Shea, some environmental item is not in fact content before it becomes exploited by some downstream process.

Mutual information, especially if we take maximal mutual information as the relevant measure, which I outline in section 7.3, significantly limits the amount of information we can attribute to the system relative to correlational information. Correlational information therefore allows for a greater range of relations to be exploited, providing the grounds for an account of content acquisition via learning. However, in section 7.5, I will argue that maximal mutual information does allow for unexploited content, since it captures all information available to the system, including information only *implicitly* available, which may not currently be processed by any given system.

7.2.2 Distinction between maxMI and infomax

Shea provides a third reason to reject mutual information. He argues that it is "an important tool for a different project, that of working out why information processing in the brain has been configured in a particular way" [Shea, 2020, p407]. He goes on to point out that "it is not the task of a theory of content to explain why a representational system is configured a certain way" [Shea, 2020, p408]. Rather, theories "of content just need to tell us what a system represents, given the way it is configured" [Shea, 2020, p408]. Shea is correct on every count.

However, I argue that it is *also* true that mutual information is an important tool for a theory of content. Specifically, in section 7.4, I will argue that the content of a representation is the item in the external environment with respect to which the representation has maximal mutual information. Finding the item which maximises mutual information is independent of an explanation of why the system is configured the way it is.

The "different project" Shea has in mind is captured by the theory known as "infomax" [Atick and Redlich, 1990]. Infomax provides an explanation of the constitution of sensory systems in terms of maximally efficient processing (elimination of redundancy). For example, the sum-difference encoding used by the ganglion cell in virtue of its opponent channel processing ensures that inputs from photoreceptors in its receptive field are decorrelated [Stone, 2018]. This has the consequence that mutual information between each part of the system is maximised, resulting in minimal energy expenditure processing redundant input.

Since maxMI and infomax are closely related, though importantly different, it is crucial to separate them in order to show that mutual information can be used in order to isolate content, not just explain the constitution of sensory systems.

MaxMI and infomax are aimed towards separate explanatory questions. Infomax answers the questions: does the system achieve maximal levels of mutual information between system components, and if so, how? Our theory answers the question: which items in the environment are explanatorily relevant for the realisation of downstream capacities?[3] We answer this, in part, by invoking mutual information. They are manifestly different projects; infomax hypothesises that two internal components are constituted

[3]Note that maxMI aims to answer *this* question, while the content it thereby isolates is used to answer *how* those capacities are realised.

so as to maximise the amount of mutual information between them; the current project aims to discover the object in the environment with which some state or structure shares maximal mutual information.

In intuitive terms, the difference in the use of mutual information can be captured by the following two questions:

1. How closely related are these two things? (maxMI)

2. How can I make these two things as closely related as possible? (infomax)

Our theory concerns (1), infomax concerns (2). Infomax is a theory which suggests there is some kind of 'target' which is approached by the system in virtue of the adjustment of internal parameters until there is a match with external parameters (i.e. properties of the statistical distribution of values). It therefore shares many properties in common with Karl Friston's free energy principle [Friston et al., 2006]. Infomax does not necessarily identify what *currently* maximises mutual information with the content of the state, nor does there appear to be any theory about what the content of a state might be within infomax[4].

In comparison, maxMI makes no claims about how the content was historically arrived at. There is no theory of the *process* which is undergone by the system in order to fix content; that is given by infomax. Instead, maxMI states that, at any given period of time, the content of a representation is the item with which the state has maximal mutual information. Infomax and maxMI are clearly compatible. Indeed, one might suggest that if the content of a state is what it currently maximises mutual information with, where

[4]See [Wiese, 2017] for discussion about whether predictive processing theories include a theory of content - he argues they do not.

this is not the most efficient item for the purposes of reducing energy expenditure, then adjusting parameters such that the system maximises mutual information with a new item which is the most efficient, captures how the system updates content in order to maximise efficiency. This process is, however, not a requirement of maxMI. Infomax is a model for a *process* within the system; maxMI is a model of the *relation* between internal and external items.

7.2.3 Summary

Shea is right that mutual information serves a different explanatory aim to correlational information. Correlational information is "the type of correlation which *natural selection* can make use of" [Shea, 2007, p240, emphasis added]. Such a notion of information is helpful for a project which aims to identify those items in the environment which best explain why a particular representational structure was reproduced. In this sense, such a project may explain *why* the organism has the content maxMI attributes to it: we might hypothesise that the content, as we understand it, correlated with some item which enabled an evolutionary advantage[5].

However, I will argue in the rest of this chapter that those interested in providing a theory of content which features in the proximal explanations of cognitive neuroscience must invoke mutual information (or a mathematically equivalent measure).

[5]See chapter two section 2.5.

7.3 Overview of maximal mutual information

Mutual information is a relation between random variables. We can express the relation in a few ways. MI is the average of the statistical interdependence of the values of two random variables given by:

$$I(X,Y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} p(x_i, y_i) log \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \tag{7.3}$$

Where $I(X,Y)$ is the mutual information between X and Y, x_{i-m_x} are the values of X, and y_{i-m_y} the values of Y.

This formulation is helpful for grasping the intuition behind the concept: a high probability of x_i and y_i together, where the marginal probabilities of x_i and of y_i are low, will yield a high amount of mutual information. The opposite case, where the probability of both is low but the marginal probabilities are high, will yield a low amount of mutual information.

Imagine the probability that I write with a pen is low, and the probability of me having ink on my hands is low, but the probability of me having ink on my hands given I have been writing with a pen is high. If you see that I have ink on my hands, you have a lot of information to the effect that I have been writing. However, if I write with a pen an average amount, but very often have ink on my hands (don't ask), seeing ink on my hands will give you very little information about whether I have been writing.

Another, simpler, formulation, explicitly stated in terms of random variables, is given by:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{7.4}$$

Where we can take X as an eRV, and Y as an iRV, $H(X)$ the entropy of the eRV and $H(Y)$ the entropy of the iRV (which defines the total amount of information available from a random variable, or the **uncertainty** of the random variable), and $H(X,Y)$ as the joint entropy of the eRV and iRV. If the random variables are independent, $H(X,Y) = H(X) + H(Y)$. Any dependencies between the two random variables will be 'squeezed out' of the joint entropy in the form of mutual information, hence equation 7.4.

Assuming we can get by with an intuitive understanding of these equations, I want to consider how we should model the relationship of *maximal* mutual information between eRVs and iRVs.

There are a range of eRVs we can take as related to a given iRV. In chapter six (section 6.5) I argued that the range of relevant eRVs is given by the range of corresponding environmental items which the C-function of the system housing the representation corresponding to the iRV is indeterminate with respect to. This generates a set of external-side random variables X_{1-n}.

Each eRV, X_{1-n}, has a set of measurable values, with a probability distribution (or density) over them, provided by their relation to the system-side invariance mechanism generating the related iRV (Y) (see chapter five section 5.4.1).

We can calculate values of mutual information with Y for each of X_{1-n} using each value of the random variable via equation 7.3. Or we can calculate entropy values and use equation 7.4[6].

[6]Or indeed any of the equations for mutual information.

7.3.1 The implicit theory: maxMI

With a list of all mutual information values between the iRV, Y, and each eRV, X_{1-n}, we can identify those eRVs with the greatest mutual information with the iRV. There may be just one eRV with the greatest MI value, or there may be several. On the current theory, if there is one random variable with maximal mutual information, this is the content of the representation corresponding to Y. If there are several, the representation evinces **natural indeterminacy**, such that each of the external items modelled by each eRV is included in what we might call, following Bergman [Bergman, 2023], the **content profile** of the representation.

The iRV itself should be modelled according to the information usable to downstream areas, based on the restriction to the C-function of the relevant subsystem. More on this in section 7.5.3.

This is what I will argue is the implicit theory of cognitive neuroscience, maxMI. Again, from chapter one (section 1.4):

> **maxMI**: E_x is the content of R iff R shares mutual information with each of a set of items, E_{1-n}, of which E_x is a member, and R and E_x have maximal mutual information relative to the rest of E_{1-n}.

Where E_x is some item external to the representation, and R is a representation, given that:

1. R must be modelled as an iRV with outcomes constrained by values usable for downstream systems.

2. E_x must be modelled as an eRV with outcomes constrained by values discriminable by sensory interfaces.

3. The set E_{1-n} must be delimited by the C-function of the subsystem containing R.

I hope that, following each previous chapter, conditions 1-3 appear relatively clear and well-motivated. Collectively, these conditions, along with the relation of maximal mutual information, provides the implicit theory of content in cognitive science. At least, that is what I will now argue. The argument primarily involves pointing out how maxMI is assumed by the methodology of cognitive neuroscience. Sections 7.4.2 to 7.4.3 argue primarily for the claim that the relation of maximal mutual information is assumed. Section 7.4.4 argues that the eRV space is delimited by the C-function of the system containing the representation.

7.4 Arguments from cognitive neuroscience

In this section I will demonstrate that cognitive neuroscience, either explicitly or implicitly, models the relation between a representation and its content in terms of maximal mutual information. I will begin by arguing that the most prominent methods for establishing representational content implicitly assume that the content is that item which shares maximal mutual information with the neural representation. I will then explain *why* cognitive neuroscience implicitly assumes maxMI: it provides a measure of the information **available** to the system, thus ensuring the explanatory value of content for answering how-questions of the sort we have identified.

7.4.1 Prediction for cognitive science

I begin with the truism about cognitive neuroscience - as expressed, for example, by Paulin and Hoffman - that practitioners "often describe the behavior of spiking neurons in terms of firing rate" [Paulin and Hoffman, 2001, p877]. That is, researchers will test the firing rates of neurons in the presence of various stimuli and make inferences about which stimuli are 'preferred' by the neural representation based on the strength of the neural response.

Why does 'preference' matter? Using firing rate is *intuitively* relevant to finding the content of a representation: if some neuron responds explosively when presented with some stimulus, we cannot help but assume that that neuron and the stimulus are somehow importantly related. But *why*?

Probability-raising must play an important part, as captured by the definition of correlational information in equation 7.1: the probability of the stimulus being present is raised when the neuron fires, providing an exploitable relation for performing various actions the success of which requires the presence of the stimulus. However, if researchers were only interested in correlational information, they would be content with finding *any* level of activity of the neuron in the presence of a stimulus. Provided the inequality in equation 7.1 were to hold, the stimulus presented to the neuron would be a candidate for the representational content of that neuron.

This is not what we typically find when we examine the cognitive neuroscience literature. Researchers are interested in much tighter statistical relations than that of correlational information. I argue that the methodology implicitly assumes maxMI. This involves two claims: first, that the methodologies involved are attempts to find the external

item with which some neural state or structure has maximal mutual information (within the range of items given by the relevant subsystem's C-function). Second, that this item is the content of the representation supported by that state or structure in the sense that it provides that item which is represented *for the system itself*. Sections 7.4.2-7.4.4 mainly deal with the first claim, section 7.5 deals with the second.

The aim of the following sections is not to provide a detailed formal demonstration that the methodological tools of cognitive neuroscience are mathematically equivalent to maximising mutual information. Such work exists, and is referenced throughout the text. The aim is merely to draw the reader's attention to such equivalences where they exist, and to make various connections between existing methodologies and maxMI explicit in relatively transparent language.

7.4.2 Spike-triggered average

The spike-triggered average (STA) of a neuron is the average stimulus which triggers the neuron to spike. The STA is used to discover both *which* elements of stimuli presented to neurons are encoded, and the *specific code* employed. As expressed by Paninski, STA is one crucial method used to answer "the most prominent" question in systems-level neuroscience: "the "what" part of the neural coding problem: what makes a given neuron in a particular part of the brain fire?" [Paninski, 2002].

In order to discover the STA of a neuron, [7] it is presented with with a range of stimuli. Stimuli are separated into those which trigger a spike, and those which do not. A novel stimulus (e.g. an image) is then composed by averaging the features of the stimuli which

[7]Note that multiple neurons can also be modelled in this way by taking a function of the spiking of each individual neuron.

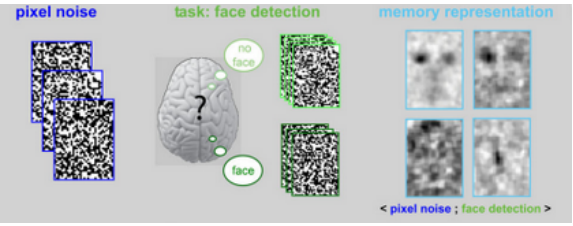


Figure 7.1: Building 'faces' from white noise from [Schyns et al., 2020].

caused the neuron to spike (see e.g. [Schwartz et al., 2006, p487]).

Imagine that some subset of a set of white noise images, presented to a neuron, activate that neuron. Each image is broken down into a number of pixels, with a value corresponding to its colour (1 = white, 0 = black). If we overlap each stimulus which triggered the neuron, we can find an average value for each pixel. This produces a new image. This image is the STA of the neuron.

Though different in important respects, the method is similar to that employed by Schyns and colleagues [Schyns et al., 2020] when probing the memory contents used for face recognition. Figure 7.1 shows images built from an average layering of white-noise images. Participants were shown white-noise images like those on the left and told (falsely) that some images contained hidden faces. They then answered 'face' or 'no face' for each separate image. Despite the random nature of the white noise, layering the white-noise images which elicited the 'face' response produces an eerily face-like image.

This suggests that participants tended to answer 'face' when the pixels were sufficiently close to a stored 'face' representation, which Schyns and colleagues took themselves to decode. This comes out when the images are averaged, even if any given single image is very unlike a face.

The STA of a cell is like this. The difference is that in Schyn et al.'s study, images

were averaged across verbal 'face' responses. The STA is the same except that the images would have been averaged across those which triggered a response from a neuron.

The average stimulus can be modelled formally, for instance as a vector comprising the various parameterised values the stimulus occupies. This allows for the construction of a mathematical model for of the specific code used by the brain to encode the stimulus. It also allows very precise specification of the external item represented.

We discussed one application of the STA method used by Chang and Tsao [Chang and Tsao, 2017] in chapter three (section 3.3.1), but finding the STA of a neuron is an increasingly common method in cognitive neuroscience (see [Pillow and Simoncelli, 2006] for an overview). According to Schwartz et al. the method has "become quite widely used experimentally" [Schwartz et al., 2006, p486] across neuroscience. The method is properly viewed as a subset of those methods which aim to discover the subspace of a stimuli which a neuron responds to, given a high-dimensional stimulus. The method allows experimenters to find the specific aspect of the stimulus which is represented. A cell may appear to respond to all and only faces, but on closer inspection may respond to a highly specific set of shape features, which could, in principle, appear on objects other than faces.

STA and maximal MI

The STA method isolates an external stimuli with which a representational state has maximal mutual information, given a few conditions. As Stone puts it[8], given some "fairly mild conditions, the average of a set of measured values is also the *least-squares estimate*

[8]The following is taken from Stone's textbook on neural information theory [Stone, 2018].

of a true (but unknown) parameter value (i.e. the true mean)" [Stone, 2018, p94]. Given some other "equally mild conditions" this provides a "*maximum likelihood estimate* of the true parameter value" [Stone, 2018, p95]. In intuitive terms, the STA method provides a good estimate (with some assumptions about the data which are likely to be met or approximated) of the probability distribution of the parameterised input.

Stone then points out that "for Gaussian variables" the maximum likelihood estimate "maximises the mutual information between those variables" [Stone, 2018, p95]. In other words, when the input distribution is Gaussian, finding the STA of the cell in response to those inputs provides you with the item in the external environment with which the cell has maximal mutual information. So, for Gaussian inputs, using the STA method implicitly identifies the item which maximises mutual information.

What about non-Gaussian inputs? There are multiple studies which do use STA for non-Gaussian inputs (see [Schwartz et al., 2006, p501]). Are these theorists *not* implicitly invoking maxMI? Not so; the item with maximal mutual information is still considered to be the content of the representation, albeit difficult to determine given the non-Guassian nature of the inputs. In such cases, theorists write that they simply need to deal with "artifacts" which can "bias" away from the "*real* subspace" [Schwartz et al., 2006, p501; emphasis added]. In other words, in such situations, any deviation from the item with which there is maximal mutual information is seen as an *error* to be *corrected*.

Indeed, in such cases, the textbook recommendation is that a more general method be used: researchers can "compare the *mutual information* between a set of filter responses and the probability of a spike occurring" to find the maximally informative dimensions (see section 7.4.3) [Schwartz et al., 2006, p502]. The drawback of this method is that it "is

significantly more complicated" - a pragmatic concern [Schwartz et al., 2006, p502]. So, it seems that purely pragmatic reasons may be stopping researchers comparing mutual information levels directly.

Whatever explicit reasons may be employed by scientists when deciding on their methodology, the fact that the STA method ends up picking out a measure which provides maximal mutual information is not a fluke. Maximal mutual information has likely been implicitly sought by previous practitioners. Pillow and Simoncelli [Pillow and Simoncelli, 2006] maintain that the methodology is approximated in previous work: "Note that 'classical' experiments can also be viewed within this framework: Characterization with dots, bars, or grating stimuli implicitly assumes that a neuron's behavior is determined by its response to a set of canonical features." [Pillow and Simoncelli, 2006, p414]. So, if STA implicitly assumes maxMI, and classical studies, such as those by Marr, implicitly assume the framework, we have good reason to expect maxMI to be implicitly assumed by a great number of cognitive neuroscience studies, and perhaps even many cognitive scientific theories developed in light of those studies.

In the next section we consider explicitly information-theoretic approaches to answering the what-question. We will see that in these approaches, maximal mutual information is either implicitly or explicitly employed.

7.4.3 Information-theoretic approaches

Information-theoretic methods involve explicit use of formal models derived from Shannon's work[9] to describe stimuli and neuronal responses to stimuli. We will cover two

[9] Almost universally Shannon's information theory, however some instances using Kolmogorov complexity can be found.

examples of information-theoretic methodology in this section - dimensionality reduction by way of maximally informative dimensions, and the use of conditional mutual information to discover represented elements of the stimulus.

First, maximally informative dimensions allow dimensionality reduction for non-Gaussian stimuli, better approximating natural stimuli and overcoming the limitations of STA. It also provides a way to reduce the dimensions of the input data when the dataset is highly correlated. For example, in natural images in which the light intensity values of one area of the image are good predictors of the light intensity values of adjacent areas. As such, it is employed by those who wish to extend the basic principles of the STA approach to a wider range of inputs[10]. It is a very general tool for a widely used methodology.

Second, conditional mutual information (CMI) as a way to discover the represented stimulus is not as ubiquitous, but is used in precisely those studies which explicitly seek to address concerns similar to Egan's, as we discussed in chapter four (section 4.4.3). For example, it is used in this way throughout the work of Schyns and colleagues in order to isolate just that element of the stimulus which is explanatorily relevant for the operation of various cognitive capacities (e.g. [Liu et al., 2022], [Schyns et al., 2020], [Ince et al., 2015]). Given this, the fact that (as I will show) CMI implicity assumes maxMI suggests that maxMI is the implicit theory of content guiding content attributions for those studies which meet Egan's criteria.

Both methods either implicitly or explicitly invoke the relation of maximal mutual information between neuronal response and that element of the stimulus which is rep-

[10] see [Pillow and Simoncelli, 2006, p415] for a discussion of the uses and practical drawbacks of the technique

resented. I will go through both in turn.

Maximally informative dimensions

Dimensionality reduction, generally, rests on the assumption that neurons in early sensory processing respond ”to a small number of stimulus features” within otherwise very feature-heavy stimuli [Sharpee et al., 2004, p3]. A picture of a house might be presented to a participant and trigger a response from one of their neurons. However, it may be unclear why the neuron is firing - it might be because of the colour of the house (the neuron may be colour-selective), the square windows of the house (the neuron may be shape-selective), the orientation of the lines describing the edge of the house (the neuron may be orientation-selective), or any number of the component parts of the image, in any possible combination - maybe the cell is responsive to orange lines oriented horizontally.

Maximally informative dimensions are those elements of the image, modelled using vectors, with which the neuronal response has maximal mutual information. As Sharpee et al., the originators of the method, describe: “we maximize the mutual information between the neural responses and projections of the stimulus onto low dimensional subspaces” [Sharpee et al., 2004, p1]. The low dimensional subspace is given by taking vectors describing the various features of the stimulus (e.g. illumination values, orientation values, colour values), then transforming those vectors into those which have minimal redundancy. This means constructing vectors whose values are composed from a number of values of the initial feature vectors, but now with minimal correlation between the features described by the new vectors. This is a similar approach to taking the principal components of a data set.

Then, levels of mutual information are found between these new vectors and the response of the neuron. The set of vectors with the maximal amount of mutual information with the neural response is considered to describe what the neuron represents.

The assumption that low-level dimensions are represented is thought to be implicitly reflected in classical work:

> the general idea of searching for low dimensional structure in high dimensional data is very old, our motivation here comes from work on the fly visual system where it was shown explicitly that patterns of action potentials in identified motion sensitive neurons are correlated with low dimensional projections of the high dimensional visual input ([de Ruyter van Steveninck et al., 1997], [Brenner et al., 2000], [Bialek and van Steveninck, 2005]). [Sharpee et al., 2004, p3]

Maximal mutual information is explicitly taken to be the relation which determines the content of the representation in the maximally informative dimensions approach. Additionally, it is thought to be implicitly assumed by classical studies in the field.

The representational content which comes out of such a view is very unlike the kind of content we might intuitively presuppose: as with previous studies we have covered, the relevant features turn out to be highly specific structural elements of the input. In particular, they are, as in Chang and Tsao's study, essentially amalgamated features based on principle components analysis. However, as I said in chapter one (section 1.6), we are not constraining ourselves to intuitions. The explanatory value of the content takes precedence.

Conditional mutual information

The conditional mutual information (CMI) approach provides an example which does not initially appear to fit with maxMI. The CMI approach does search for the aspect of the stimulus which is represented, but does not straightforwardly do so by finding the item (or aspect thereof) with which neuronal firing has maximal mutual information.

Instead, CMI, as we saw in chapter four (section 4.4.3), involves finding the mutual information between two random variables conditioned on a third random variable, given by:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{7.5}$$

As we also saw in chapter four, CMI can tell us whether the variable Z partly (or perhaps fully) accounts for the interaction between X and Y. So, CMI appears to look for at least three variables, not two, in order to determine what is represented. Rather than find two variables, an iRV and an eRV, which share maximal mutual information, it looks for a third eRV upon which to condition the response between the iRV and eRV, then determines whether this second eRV is controlling the iRV.

Does the CMI approach use a different implicit theory from maxMI? It does not. If we look at how CMI is used in studies which attempt to find what is represented, it becomes clear that CMI does not deviate from the implicit theory of maxMI. In fact, the CMI approach provides just another route to finding the variable with which the iRV has maximal mutual information. This is clear once we reconceptualise the relation between the variables. I will provide a theoretical reconceptualisaton before providing examples

of how this works in practice.

CMI can be used to find the eRV, Z with which the iRV, X, has maximal mutual information if we can discover an eRV, Z, for which $I(X; Y_{1-n}|Z) = 0$ for a range of eRVs given by Y_{1-n}, for all values of n, within the given stimulus space (defined by the C-function of the relevant subsystem). This is just to say that Z has maximal mutual information with X. In such a case, there is no other eRV the addition of which to Z reduces the uncertainty of X any more than the uncertainty reduction achieved by Z alone. $H(X|Y_n, Z)$ is no larger or smaller than $H(X|Z)$ - so, Y_n is effectively 'screened off' from any statistical relationship with X.

If we found that $I(X; Y_n|Z) \neq 0$ for some n, we would have located some eRV which accounts for some aspect of X over and above Z. In this case, we can work to discover some combined feature which can be re-modelled as its own eRV which takes as its values some combination of values of Z and Y_n. Or, we may model a number of complex features as a set of eRVs with respect to which the iRV has maximal mutual information taken as a whole, which would therefore define the content profile of the representation.

This is not the only way to discover what maximises mutual information using CMI, but it may be the most intuitive example. Other approaches use various techniques, using various aspects of CMI in order to generate algorithms for dimensionality reduction, which involves finding those features of the stimulus which are represented by the neural response (e.g. [Souza et al., 2022], [Liang et al., 2019]).

On this interpretation, we can use the CMI approach to find increasingly complex assemblies of eRVs, or a new eRV constructed from values of other eRVs, with respect to which mutual information with the iRV is maximised, over and above the level of mutual

information shared between the iRV and just one of the original eRVs. An example of this approach is found in Liu et al. [Liu et al., 2022]. Liu et al. found, using CMI, that certain facial expressions 'multiplex' both emotion category (e.g. sadness, happiness) and dimension (e.g. intensity, valence). CMI revealed that some facial expression components measured in AUs (action units) carry information about both the emotion category and its strength, (e.g. a furrowed brow represents both disgust and the intensity of the disgust). The experimenters constructed new feature sets which collectively have maximal mutual information with the relevant AU[11]. In other words, their methodology is based on the assumption that the standalone represented categories are found by discovering those categories and dimensions, separately, which maximise mutual information with the response of participants. The combined, or 'multiplexed' features are those which together fully reduce the uncertainty of the response, and therefore which provide the complex feature which maximises mutual information with the response.

In essence, using CMI uncovers more complex eRVs by way of discovering relations between environmental items and responses, providing a new environmental item with respect to which the response maximises mutual information, exhausting the possible stimulus space by combining existing, independent elements. It can be used to overcome limitations of taking mutual information between pre-determined environmental items and responses, and instead comparing levels of mutual information when conditioning on other eRVs.

Rather than being inconsistent with maxMI, the CMI method instead provides a *way*

[11]It should be noted that this study flips the content and representation: in this study, facial expressions are taken as representations and emotion categories and intensities of the perceiver of the face are taken as their contents. The experimenters are interested in what the face encodes in this instance, rather than what a neuron encodes. However, the analysis is otherwise the same.

to discover the relevant eRV, or set of eRVs, with respect to which the iRV has maximal mutual information. In sum, we should conceive of CMI as a way of searching for an eRV (or set therof) with maximal mutual information with the iRV. As a gold standard method of answering the concerns of Egan and de-Wit et al., the fact that CMI presupposes maxMI provides good evidence that maxMI is the implicit theory used by those regions of cognitive neuroscience in which content forms part of the theory proper.

7.4.4 maxMI: eRV space restricted by C-function

So far, I have argued that the above methodologies assume that the relation of maximal mutual information is the relevant content-determining relation. However, maxMI includes a commitment to not only the relation of maximal mutual information, but the relation of maximal mutual information *within* the range of possible eRVs determined by the C-function of the subsystem containing the representation. Is this supported by the methodology of cognitive neuroscience? The argument here is thankfully quite short.

As I argued in chapter six (section 6.4), functions in cognitive neuroscience *are* C-functions. So, if any function constrains content attribution in cognitive neuroscience it is C-function rather than W-function. But why think that C-function constrains content attribution?

First, studies almost always begin from a hypothesis about the content of a representation based on existing theories about the function of the representation's subsystem (isolated typically in terms of cortical region). For example, Chang and Tsao initiate their 2017 study by observing that "A central challenge of visual neuroscience is to understand how the brain represents the identity of a complex object. This process is thought to hap-

pen in inferotemporal (IT) cortex" [Chang and Tsao, 2017, p1013]. So, their search for the specific content of cells within IT is restricted to those items which they think are able to support the function of IT.

Generally, there is back and forth. Functions are initially attributed to cortical regions based on the general set of stimuli found to trigger neurons in those regions. More complex theories of function develop, such as those initially proposed by Marr for the visual system [Marr, 2010]. Then, based on these theories, more specific searches for content are enabled. At this point, content can be isolated which plays a genuine explanatory role in how cognitive capacities are enabled.

Second, studies which look for content typically involve relating that content to downstream cognitive capacities. For example, DiCarlo et al. note that object recognition must be evaluated relative to "defined tasks that can be measured in behavior, neuronal populations, and bio-inspired algorithms" [DiCarlo et al., 2012, p429]. Content is discovered relative to its role in driving behaviours which correspond to tasks. C-functions capture this relationship, since they are ascribed relative to the role of some subsystem in an explanation of the performance of some downstream capacity.

If the above is correct, maxMI is the implicit theory of content in cognitive neuroscience.

7.4.5 Scope of maxMI

The above statement needs a little nuance. While many studies do implicitly rely on C-functions, this does not always translate into practice in the right way. The now-familiar challenge from de-Wit et al. [de Wit et al., 2016] attests to this. Even if researchers look

for the relationship of maximal mutual information, and implicitly *recognise* the role of C-functions, this will not always result in theorists limiting content attributions by decodable information. That is, C-function ascription will not place the right kind of limits on content attribution.

We have seen in previous chapters how researchers such as Schyns and colleagues explicitly address this issue. Researchers such as Chang and Tsao at least include a *hypothesis* about what can be decoded by the system. However, it is unlikely that every researcher makes this restriction. Instead, they will take the neuron or neural assembly *as such* and find what it maximises mutual information with, rather than taking as an iRV just that element of the neuron which is readable by downstream systems.

In many cases, this will likely be an unproblematic simplification. This is Neander's view:

> We can (to a first approximation) trust that a creature's sensory-perceptual systems have been adapted to provide information that its other cognitive systems can use, and that a creature's other cognitive systems will have been adapted to exploit the information that its sensory-perceptual systems can provide. [Neander, 2017b, p144]

Neander justifies this by noting that information which *could* be extracted from a neuron probably will be, since the neuron providing that information to begin with "is not cheap; it is costly" [Neander, 2017b, p144]. However, we can just as well respond that further processing that information is also very expensive, energy-wise, so we cannot just assume that all information-transmission inside the system will be non-lossy.

So, what to make of those studies which currently do not adhere to the strictures of maxMI? On the extreme end of the spectrum, where *no* implicit theory is governing content-attribution, content will be a gloss. I think we will tend to find this gloss in the opening and closing remarks of a study, in which the scientists are attempting to broaden the scope of their findings in an intuitive way, as a prelude to more theoretical work. As content attributions move along the spectrum, and start to mirror the requirements of maxMI, we should see them as helpful first passes, which oversimplify, but which can feed into more nuanced future studies by suggesting paths of investigation.

7.5 Availability

I close this chapter by considering how maxMI secures content attributions which are explanatorily relevant for answering how-questions in cognitive neuroscience. I argue that maxMI isolates content about which the system has **available** information. Informally, available information is the information which the system itself can retrieve from its own states. It defines an upper limit on the complexity of the content which the system is able to represent. This distinguishes maxMI from etiological teleosemantic views, in which content is limited by *external* correlations, but faces no limit based on internal processing capacity.

Availability of information secures the explanatory value of maxMI content in virtue of isolating content which, when modelled as an eRV, possesses values all of which are able to undergo processing by the system itself. This allows for a mechanistic explanation of how the external item is processed and used for cognitive tasks.

I begin by elucidating availability as understood in the neuroscience literature. I then show how maxMI provides content about which the system has available information. I then argue that availability, hence maxMI, secures explanatorily relevant content.

7.5.1 Implicit and explicit availability

Discussions around availability centre on the distinction between information which is either **implicitly** or **explicitly** available to the system.

Two typical definitions of explicit availability are given by Kriegeskorte and Diedrichsen [Kriegeskorte and Diedrichsen, 2019] and Kirsh [Kirsh, 2006].

Kriegeskorte and Diedrichsen define an "explicit representation" as "a representation of content in a format that enables it to be decoded in a single step by biological neurons" [Kriegeskorte and Diedrichsen, 2019, p411]. "A single step" means no further mathematical operations have to be applied for information about the representation's content to be used in some further downstream process. A typical case is one in which the representation contains "information accessible to *linear decoders*" [Kriegeskorte and Diedrichsen, 2019, p411; emphasis added]. Linear decoding, in terms of the physical implementation of the decoder, involves a step change in firing rate of an input neuron that is related linearly to a step change in the post-synaptic neuron.

Similarly, Kirsh outlines the distinction between implicit and explicit availability in terms of "the computational effort required to extract, use, or interpret the information encoded in a representation" [Kirsh, 2006, p479]. The emphasis on the extraction or use of the information is reflected by Kriegeskorte and Diedrichsen: they stipulate that a "decoding model" should "take brain responses as input and predict downstream brain or

behavioral responses" [Kriegeskorte and Diedrichsen, 2019, p409], since decoding of inputs requires downstream uptake of the encoded information for use in further cognitive capacities.

Unlike Kriegeskorte and Diedrichsen, Kirsh views explicit and implicit availability as a spectrum in which the "computational complexity of the process of interpretation determines where on the continuum of explicit to implicit a given representation lies" [Kirsh, 2006, p479]. In order to determine how explicit or implicit the information contained in a representation is, Kirsh points out that we must assume that "it is possible to use techniques of computational complexity theory to measure the computational effort involved in recovering the information" [Kirsh, 2006, p480]. This ensures that explicitness is well-defined and tractably quantifiable.

We need not be concerned, at this stage, with arbitrating between Kriegeskorte and Diedrichsen's binary linear/non-linear distinction between explicit and implicit[12], and Kirsh's continuum-based computational complexity approach. Instead, I wish to focus on an implication common to both characterisations of availability: information is available to a system, implicitly or explicitly, only if there is some downstream process which can **decode** the values of the input *at all*. So, information about some random variable is available to a system only if some mathematical function can be performed, by the system, over the values of that random variable.

A consequence of this definition, Kirsh highlights, is that it makes available information system-relative. Whether downstream areas are able to decode the input will

[12]Kriegeskorte and Diedrichsen also acknowledge 'degrees' of explicit availability - from information linearly decodable by to a single, immediately downstream neuron to information linearly decodable by a set of long-range, distributed neurons [Kriegeskorte and Diedrichsen, 2019, p423].

vary according to the computational capacities of those downstream areas. Commenting on sophisticated representational contents, Kirsh notes that "individual capacities for memory, learning and other cognitive skills can affect how explicit a representation is" [Kirsh, 2006, p480]. Kirsh notes that if there are "no connections or accessing procedures that can reliably make use of the information" in a representation, then information about its content is not available. Differences in neural connectivity and the existence or lack of complex non-linear decoders can change which information is available to the system. To anticipate, this means that on the maxMI view, the content of a given representation can vary depending on the properties of the system housing the representation.

7.5.2 Availability and explanatory value

Throughout the thesis, we have accepted Egan's criteria for inclusion in a theory proper in cognitive neuroscience. According to Egan's characterisation, the mathematical function computed by the system is part of the theory proper. We can now see the specific role it plays; information is available to downstream systems only if there is a mathematical function performed which translates values of the input into values of the output. So, the mathematical function computed enables us to determine the available information, and how that information is encoded and decoded.

The theory proper also includes environmental items which the externalised cognitive capacity has an explanatory relation to, or so I have argued throughout the thesis. I argue that, to be explanatory of external-directed capacities, there must be an information-processing chain which runs from the item in the external environment, and is such that we can trace operations over eRVs, through iRVs, all the way to the mecha-

nisms controlling the eventual output. In order for information about an external item to be available, there must be a mathematical function computed which performs part of the explanation for how the information related to the input is transferred throughout the cognitive system. In order for some information *about* some input to be available to the system itself, either implicitly or explicitly, we must be able to specify precisely which input values are taken in from the environment, into the system for further processing. We must trace the information flows from the content, through each part of the internal sender-receiver relay, to the terminal receiver.

I argue for this claim in a familiar way: it is implicitly assumed by neursocientific practice. In systems neuroscience the what- and how-questions are related: researchers want to know *both* what is processed and how it is processed - with the tacit assumption being that the information over which operations are performed can be traced back to the input, which terminates in the external environment. More specifically, researchers wish to know "what information is discarded in the neural code, and what features are most important" [Paninski, 2002, p1]. That is: from the input, what is retained for further processing?

This is most clearly the case in psychophysics, which is why we chose to focus on the discipline in chapter four (section 4.4.2). In psychophysics, information flows are traced from sensory interfaces to higher cognitive areas (see section 4.4.4). A redundancy measure, RED, is used to trace the information involved in perceptual decision-making, beginning from the stimulus, tracing the informational path to sensory interfaces, then to downstream regions of the cognitive system. This tacitly assumes that the relevant information is available to the system. The values of the input are taken to be computed

over by the rest of the system; the assumption behind the RED measure, for example, is that the values of the stimulus which share mutual information with a response are redundant with respect to the values that the behavioural decision shares with the response. The system receives no more information from the stimulus than it does about the upstream response: all the relevant information about the stimulus is available in the response.

The general principle which explains why these disciplines assume availability is provided by Kirsh: an assumption across cognitive science is that "a representation is a well-defined state, structure or process, in a causal system; that it encodes a specifiable informational content that can be harnessed by the causal system of which it is a part" [Kirsh, 2006, p479-80].

I attempt to remain non-committal on the question of the particular kind of explanation offered in cognitive science, so will not comment on whether scientific explanations are causal in any strong sense. However, if there exists an *assumption* that representational content features in a causal explanation, there must be an implicitly assumed causal *mechanism*. Whatever one's view on causality, there must surely be some way of transmitting causal influence (such as Salmon's causal marks [Salmon, 1984]). If some item in the world is to be causally relevant for internal processing, we must hypothesise *how* that causal influence is affected. This must be the case for the informational content which is harnessed by the system. For any value of the eRV we take to have some causal influence on the system, we must be able to show how some part of the system relevantly corresponds to that value, such that that part of the system can causally interact in such a way as if the value of the eRV itself were providing direct stimulation.

This captures the philosophical description of representations as "stand-ins" for external items (e.g. [Cao and Warren, 2023]) This requires that the information be available to the system - that we can trace a processing link back to the original external value.

If this is correct, the explanatory value of content is secured only if information about that content is available to the system itself. In the next section I show how maxMI delivers contents about which the system has available information.

7.5.3 maxMI and availability

The relation of maximal mutual information between an iRV and eRV provides, given some conditions spelled out below, a measure of the *implicit* information about some environmental item conveyed by a representation.

According to Kriegeskorte and Diedrichsen, the total available information or "all encoded information" about a stimulus feature is given by the "mutual information between any stimulus feature (graded or categorical property of each stimulus) and the response pattern." [Kriegeskorte and Diedrichsen, 2019, p410]. Mutual information provides the "total information that the representation contains about the stimuli"
[Kriegeskorte and Diedrichsen, 2019, p418]. Looking beyond the information readable by linear decoders, mutual information supplies the "information about any other features that may be present in the code, as well as information that would require a more sophisticated (e.g., nonlinear) decoder" [Kriegeskorte and Diedrichsen, 2019, p419].

If we define some iRV, and measure the mutual information between that iRV and our eRV, we will find how much information the eRV contains about the iRV. However, the amount of information the eRV contains may be quite low. There may be many values of

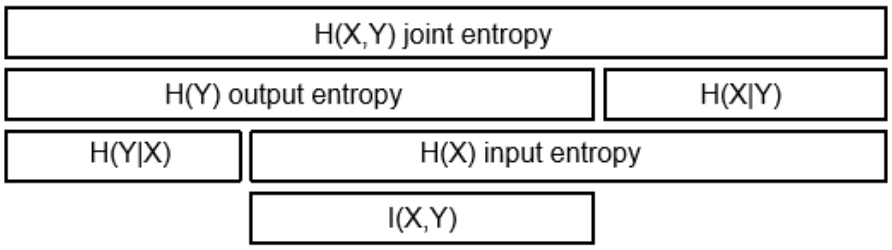


Figure 7.2: Visual representation of mutual information

the eRV which are not processed by the system. In figure 7.2 we see an example in which an eRV, X, has mutual information with an iRV, Y, but in which X has significantly higher entropy than X, denoted by $H(X|Y)$ - the entropy (or uncertainty) left in X given a value of Y.

Another formula for mutual information is given by:

$$I(X,Y) = H(X) - H(X|Y) \tag{7.6}$$

However, minimising $H(X|Y)$ - the uncertainty left in X given a value of Y - maximises mutual information. If we find a value of mutual information between two RVs for which there is a large amount of residual uncertainty about the eRV, we can define a new eRV (for example, using CMI - section 7.4.3) for which $H(X|Y)$ is smaller and mutual information is therefore larger. Iterating this process, maximising mutual information, ultimately provides us with an eRV in which uncertainty is minimised. Uncertainty will never be entirely reduced - see below - but maximising mutual information in this way provides us with an external item no value of which (outside of noise - again, see below) fails to be processed by the system itself.

Noise

In reality, there will always be some amount of **noise** between an eRV and iRV. Some value is due to noise if that value *does not* reduce uncertainty about a source. Noise models phenomena such as inefficiencies in encoding (for example, redundancy) or physical properties of the system (for example, voltage degradation over axonal channels).

Inefficiencies can be measured. As Borst and Theunissen write, since "$H(R)$ represents the maximal information that could be carried by the neuron being studied, comparing $H(R|S)$ to $H(R)$ gives an estimate of the neural code's efficiency"
[Borst and Theunissen, 1999, p949]. $H(R)$ provides the output entropy of the neuron, while $H(R|S)$ provides the conditional entropy of the neuron given the stimulus - it reflects how much is left uncertain about the neural output once everything is known about the stimulus. This is the noise due to inefficiency. However, note that this variety of noise does not suggest that there are values of the eRV which are not computed by the system. Rather, it means that there are values within the system which are *redundant* with respect to the values strictly needed to encode the content.

Noise due to physical properties of the system such as voltage degradation are more troubling. For this kind of noise, there typically *are* values of the eRV which are *not* operated over by downstream systems. Some values are lost during processing. How do we square this with the explanatory value of content according to maxMI? Note that noise, on this account, is given by features of the occurrent functioning of the state, but maximal mutual information provides values which *can* be decoded by downstream systems, even if they, on a given occasion, are not. Borst and Theunissen compare the relation provided by mutual information as supplying that which "an ideal observer"

could use to "discriminate between the stimulus conditions" [Borst and Theunissen, 1999, p948].

If everything in the system were running perfectly, there would be no noise of this variety, and every value of the stimulus would be transferred throughout the cognitive system. However, in nature, things rarely work so efficiently. Even at the basic photoreceptor level, random heat fluctuations can lead to the hyperpolarisation of a cell, introducing noise. This does not preclude the explanatory value of a model which explains the performance of photoreceptors in terms of interactions with photons - the operation of the cell under ideal conditions. The same is true for content picked out by maxMI.

It is true that maxMI would not isolate content about which the system has implicitly available information if we make no restriction on what we model as the iRV. Consider again an implication of the computational definition of availability: what is available is relative to the downstream systems and what they can decode. If we neglect this when we calculate the mutual information between an iRV and an eRV, we will, in all likelihood, come up with an eRV which has values which are *not* decodable by the system. What *we* can retrieve from neural activity, given everything *we* know, is likely to outstrip with the system itself can retrieve from that activity. This is a reiteration of the worry expressed by de-Wit et al. - we would be discovering the available information taking the "experimenter-as-receiver" rather than the "cortex-as-receiver" [de Wit et al., 2016, p1415].

Usable information

In order to overcome this we need to restrict the iRV such that the values we include are picked up by the system. Imagine that we have a neuron which spikes 0.1 times a second, but the immediately downstream neuron only responds to every other spike. In this case, if no other neuron is responsive to the missed spikes, our iRV should model as values only every other spike of the neuron.

This is to limit the iRV to the **usable** information. As Martin Elliffe writes, "not all measurable information may be usable information" [Elliffe, 2000, p180] since we need to be clear about "the precise nature of that which is being measured - information is only usable information when the form of the decoder is appropriate" [Elliffe, 2000, p198]. We must understand which values of the neural code are picked up by any downstream systems - ideally, with a theory about the format of the code and the decoding capacities of the rest of the system.

Once we have made the restriction that the iRV be modelled according to those values which are usable, maximal mutual information between that iRV and some eRV provides us with the values of the eRV which are implicitly available in the neural representation, thus, I argue, providing us with a model of the content of that representation.

Usable information is a notion which can be given independently of the question of what that information is *about*. Usable information is that information, considered purely in terms of the symbols of the code, which can be processed - mathematically operated over - by downstream areas. *Available* information goes beyond what is usable: it isolates the *content* of the usable information. Available information tells us which environmental items are those which are relevantly related to those symbols which receive mathematical

processing by the brain.

maxMI restricts the iRV to usable information given the use of C-functions as a way of restricting the space of possible eRVs. Recall that a C-function is defined as follows:

> x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s's capacity to ψ just in case x is capable of ϕ-ing in s and A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to ϕ in s [Cummins, 1975, p762]

I claim that in order for an account which uses representation to "appropriately and adequately account" for a capacity, where this involves the fact that the representation is capable of *doing something* (ϕ-ing) within the system, the representation ought to have some identifiable *means* of doing that thing. There needs to be some way in which the representation interacts with the downstream structures which enact the capacity.

To summarise, maxMI isolates content which is explanatorily relevant for the system. It limits the iRV to those values which are usable by downstream systems (without yet taking into account what that information is about). It limits the range of possible eRVs to those which are within the indeterminacy profile of the C-function of the relevant subsystem. It then picks out an eRV with which the system has maximal mutual information which, given the iRV restriction, ensures that all values of the eRV are available to the system itself, providing the grounds for a mechanistic explanation of the contribution of the content to the relevant cognitive capacity.

7.6 Conclusion

Cognitive neuroscience attempts to find items which are explanatory relative to how a system enacts cognitive capacities. They implicitly rely on maxMI, isolating that content with respect to which the representational state maximises mutual information, given restrictions based on the C-function of the system housing the representation. Not all scientific theories rely on the entirety of maxMI, but those which do completely fulfill the requirements set out by Egan for content to be included in the theory proper. Those which don't either use content as a helpful gloss, or approximate genuinely explanatory content - for example, classical studies which use maximal responsivity of a cell which approximate the item with which mutual information is maximised.

I claim that neuroscientists implicitly rely on maxMI because it provides the item about which the system has available information. By maximising mutual information, we are able to find the item relative to which the system minimises noise, allowing - given some relatively idealised circumstances, as is common to all scientific practice - all values of the eRV to be processed by the system. In this way, maxMI allows contents to be isolated which are contents *for the system itself.* A change in the content leads to a change for the system itself. If some *per mirabile* change occurs in the environment, resulting in a change in the item modelled by the eRV, the system itself will notice; any outcome change makes a difference to the system, since all values are processed and used to enact some cognitive capacity. So, we have found an implicit theory of content in cognitive science which meets Egan's criteria for inclusion in the theory proper.

Theories which set out to answer ultimate why-questions do not abide by maxMI, and they need not. However, maxMI can be of use to such theorists. I hope that it can

isolate items - those which I am calling content - which further environmental items are themselves correlated with, those which can be used to answer why-questions. Some representations may maximise mutual information with dimming light; if so, we may explain *why* we have such representations by invoking the correlation between dimming light and night-time.

I hope that by making the implicit theory explicit, cognitive scientists and philosophers alike will have a tool to clarify which content attributions serve which projects. I also hope that we will be able to enrich projects in which content features by providing a fully-costed set of justifications for content attribution.

Chapter 8

Conclusion

8.1 Introduction

In this thesis, I set out to answer the following questions:

1. Is there an implicit theory of content in cognitive science?
2. How can we discover the implicit theory of content?
3. What is the implicit theory of content?
    i. Which type of information link is relevant for content determination?
    ii. Which type of function is relevant for content determination?

To attempt to answer these questions, first, in chapter two, I took guidance from existing naturalistic theories of content. I provided a brief history of informational teleosemantics and presented some of the key concepts developed which would, hopefully, aid

our search for the implicit theory of content in cognitive science, if such a thing was to be found. From Dretske, we saw the power of Shannon's information theory to model the relation between representation and content. From Millikan we saw the importance of functions to provide content with an impact for the system itself. From Neander we took response and information-transmitting functions, as well as saw the importance of clarifying and distinguishing possible explanatory projects. From contemporary authors we took much, such as clarity on the role of information theory, insights into how to approach proximal explanations, and an understanding of the role of the receiver in our model.

In chapter three, I argued that there is, in fact, an implicit theory of content in certain regions of cognitive science. I introduced a study on face recognition by Chang and Tsao [Chang and Tsao, 2017] in which content features in the theory proper. So, it appeared that we should be optimistic about a positive answer to question (1). However, we took seriously Egan's criteria on content being part of the theory proper, and so narrowed our search to those studies which treat content as essential and determined by a naturalistic, sufficiently determinate, relation.

In that chapter, I also set out three principles to guide us towards studies in which we might find the implicit theory of content. I suggested that we must, first, focus on studies which posit representations which serve a function for an externally terminating cognitive capacity. Second, focus on studies which provide a hypothesis about what the system itself can decode, or otherwise access. Third, focus on studies which describe content using technical terminology. So, we began to answer question (2).

In further pursuit of the implicit theory, I introduced, in chapter four, the background

theoretical framework of content attribution in the relevant regions of cognitive science. I argued that Shannon's mathematical framework of information theory is used in contemporary studies which explicitly seek to address the kinds of worries raised by Egan. I raised the worry presented by de-Wit et al. [de Wit et al., 2016] that many studies only take the "experimenter-as-receiver" rather than the "cortex-as-receiver". However, I outlined how information theory, which is *not* limited to simple correlations, is used to pinpoint precisely the element of the environment which is picked up by the system *and* used by downstream systems to perform a cognitive task. We also saw in chapter four that the information link of maximal mutual information appears to be assumed by practitioners.

In chapter five, I raised Shannon's warning. If information theory is the background theoretical framework of content attribution, we need to be clear and careful about how to apply information theory to the cognitive system. Indeed, the application of information theory presents constraints on what we can model. We must be able to specify aspects of the world to model as random variables. Moreover, we must do so in a way which is relevant to content determination. We need to find measurable outcome values with corresponding probabilities, summing to unity. I argued that we should model the iRV according to those elements of neural firing which are causally detectable by downstream systems. I also argued that we should model the eRV according to those elements of the environment which are causally detectable by sensory interfaces. I provided some detail on precisely how we might do this by using response profiles, receptive fields, and invariance mechanisms.

In chapter six, I introduced a further worry specific to maxMI. If we are to compare

levels of mutual information, we need to be able to specify a relevant reference class against which to make comparisons. I argued that the function of the system containing the iRV provides the requisite limitation. Moreover, I argued that we should consider the function as Cummins' does - which I called the C-function. Chapter five encouraged us to limit our search for an implicit theory to cognitive neuroscience. The relevant function ascriptions, due to the possibility of extreme pluripotency, are made without recourse to learning. Functions are ascribed, rather, consistently with Cummins' analysis. As such, we should use C-functions. I then spelled out how C-functions are able to limit the range of eRVs to those which can, in principle, explain the cognitive capacity under consideration. C-functions limit eRVs to those which can, in principle, be used by the system itself.

Finally, in chapter seven, I outlined the central argument for maxMI. The relation of maximal mutual information is implicitly taken to isolate the content of a representation, since it is implicit in the methods used in cognitive neuroscience to discover what a representation represents. C-functions determine the relevant range of eRVs: it is within this domain that neuroscientists search for content. Given the restrictions on the iRV and eRV we set out in previous chapters, maximal mutual information provides the item in the environment about which the system has available information. Noise is reduced to the extent that, given certain idealised conditions, the system processes each value of the content. So, I argued, we discovered the implicit theory of content in cognitive science which isolates contents which feature in the theory proper - contents a change in which results in a change for the system itself.

In summary:

1. Yes, in certain regions of cognitive neuroscience.

2. By attending to the methodologies of those regions of cognitive science which address the kinds of concerns raised by theorists such as Egan and de-Wit et al..

3. The implicit theory of content is maxMI.

    i. The information link of maximal mutual information is relevant for content determination.

    ii. C-functions are the relevant kind of function for content determination.

8.2 Some implications of maxMI

In this section, I briefly explore some implications and applications of maxMI.

8.2.1 Representation for the system itself

As stressed at various points throughout the thesis, maxMI isolates content which makes a difference for the system itself. This was specified as content a change in which results in a change for the system. Given that, under certain conditions, each value of the content is processed by the system, a change in the values of the eRV will lead to a change in processing, with implications for the cognitive capacity enabled by that processing. If our putative $\langle shape \rangle$ representations in AM IT suddenly maximise mutual information with tree branches, thus changing the content to $\langle branch \rangle$ (given that 'branch' is operationalised), we may start greeting trees as old friends.

In this section I wish to put representation for the system itself in more intuitive terms, indicating, somewhat vaguely, the wider value I believe maxMI has. Representation for the system itself captures what is going on for an organism (for example). It tells us how the organism itself views the world around it - the information the organism takes in and processes. *We* may look at an ant and know that it is responding to the sun, since *we* can represent the sun (necessarily - the word 'sun' picks out our own way of viewing the world). The ant presumably, with its limited processing capacity relative to our own, picks up far less information than is required to represent the sun as such. It probably represents some far more proximal input, which makes sense of the fact that we can observe the very same behaviours when we place a giant UV lightbulb near the ant.

Why the ant represents this proximal input is very likely to be explained by the fact that the direction of UV light in the ant's environment typically corresponds to the location of the sun, and following the sun is adaptive for the ant. This is a type of (high-church) content relevant for ultimate explanations, but should be distinguished from the type of (low-church) content relevant for proximal explanations.

I suggest that maxMI is useful for non-human animal psychology and comparative psychology. Since maxMI picks out representation for the system itself, we stand to learn a lot about the differences between animals in terms of the content of their representations. It should also be useful for studies in animal communication. Using maxMI, we can grasp the information available to the organism which it is able to transmit to other organisms who are able to decode it. Theories of animal communication which build on information theory, such as Mitch Green's signalling view (e.g. [Green, 2017]), are a natural fit.

8.2.2 Empirical considerations

A benefit of maxMI, and the distinctions made between proximal and ultimate projects, is that content attribution becomes an empirical project, rather than a purely philosophical one. We cannot attribute contents from the armchair. Generally, intuitions play little role in content determination. Rather, content is specified by technical terminology which picks out phenomena in the environment in ways we may struggle to conceptualise intuitively. For example, taking the 6 most informative dimensions of shape (as in Chang and Tsao [Chang and Tsao, 2017]) involves amalgamating various shape properties into novel features for which we have no existing concepts.

maxMI specifies the kind of empirical data we must have about organisms and the environment in order to make such content attributions. We need to know about internal connectivity - which subsystems connect to which other subsystems both upstream and downstream. We need to know about decoding capacity, processing limitations, discriminatory limitations and capacities, as well as facts about the environment used to determine the outcomes and probability distributions of eRVs.

With respect to the empirical facts required, I take it that maxMI straddles, and in some way reconciles, internalism and externalism about representation. It is an externalist view in some respects - contents are external items, and changes in facts about the environment alone can lead to changes in representational content. It is an internalist view in other respects - contents are limited by discriminatory capacities and internal processing constraints. This is because maxMI is aimed at proximal explanations, but proximal explanations which inherently involve the external environment. It is aimed at explaining how a system performs cognitive capacities which involve things external to

it - such as faces or objects. But being aimed at *how* a system does this requires attending to what the system is capable of doing and its limitations.

8.2.3 The value of noise

Dretske requires that informational content is specified by that item with which a representation has conditional information of unity (see section 2.2.1). However, maxMI tolerates noise between representation and content. What *maximises* mutual information is what is relevant, but that is consistent with some things activating the representation which are not the item which is its content. As such, this allows activation of the representation by internal systems, providing use of the content *offline*. In other words, it allows that the representation can be *decoupled* from its input.

For example, top-down processing may activate lower-level sensory representations for recruitment in linguistic interpretation. Authors such as Lawrence Barsalou (e.g. [Barsalou, 1999]) and Diane Pecher (e.g. [Pecher et al., 2003]) provide evidence that this is the case, and maxMI is consistent with their results. If we include this top-down activation in the mutual information profile of the representation, provided that it does not maximise mutual information with the representation, it does not risk the content of the representation changing from its external item.

8.3 Scalability of maxMI

I said in the introduction that maxMI is a modest theory of representation in Peter Godfrey-Smith's sense [Godfrey-Smith, 1998]: it aims to capture the content of relatively low-

level representations, those Neander characterises as "nonconceptual or preconceptual sensory-perceptual representations (perhaps together with a relatively small set of core concepts)" [Neander, 2017b, p10].

However modest I might try to be, I struggle to constrain my ambitions. I think that maxMI might scale, and would like, in the safety of the conclusion, to speculate a little.

According to maxMI, content is limited by decodability and processing constraints. So, as the complexity of downstream systems increases, along with the information storage of those systems, facilitating greater information retrieval from upstream areas, the limitations on content are increasingly lifted. The higher the limit on the information available to the system, the greater the entropy of the eRV with maximal mutual information with a given iRV, and so the more complex the content.

Content is also limited by the discriminatory capacities of upstream sensory systems. So, the greater the complexity of sensory systems, and the wider the range of upstream inputs to a given representation - including multimodal inputs - the more complex the discrimination profile and, again, the more complex the content can be.

If it is possible for a representation to be recruited by multiple downstream areas for a given cognitive capacity, with a corresponding increase in the amount of information which can be decoded from that representation, along with greater input variability, then contents could be incredibly complex and sophisticated - perhaps reaching the level of richness possessed by concepts.

However, many high-level representations may not feature in the same kind of explanations as lower-level representations. For example, while $\langle shape \rangle$ representations in AM IT feed directly into face recognition a high-level conceptual representation, such

as the concept CAT specifying $\langle cat \rangle$, may not obviously feed directly into any specific cognitive capacity. As such, conceptual contents may not need to be limited by the requirement that all their values be processed by the system. This requirement ensures that a mechanistic explanation of how the capacity is enacted can be given. If concepts do not need to meet this requirement, it may mean that they could represent their contents far more noisily. So, while CAT may maximise mutual information with cats, thus having the content $\langle cat \rangle$, it could be that there is a lot of information about cats that the system does not have available to it.

However, this would not mean that this content would not feature in any explanations at all. We may explain the searching, investigating, questioning, explaining nature of human beings, for example, in virtue of the fact that we are attempting to gain more available information about the content of our conceptual representations. We would explain this by pointing out that by *our own lights* we do not know everything about the things we can represent. In fact, we may have metacognitive processes which monitor the level of noise of our concepts and motivate us to reduce the noise to levels which enable us to use our conceptual representations as precisely as we can lower-level representations. Maybe.

8.4 Future areas of investigation

In this section I list some questions which are not addressed in the thesis, but which are of interest to future investigation.

8.4.1 Types of representation

Does maxMI capture all types of representation? We have looked almost exclusively at response-based neurons, but is maxMI of use to characterise the content of representations such as those recruited in memory, thought to be stored in either engrams or the chemical structure of cells (see e.g. [Gallistel, 2017])?

8.4.2 More indeterminacy

How should we respond to some of the varieties of indeterminacy not covered in this thesis already (in sections 3.4, 6.6.1, 7.3.1)? For example, what about the problem of disjunction? Is it the case that an iRV will always maximise mutual information with a disjunctive set of eRVs relative to any given single eRV? That is, will it always be the case that:

$$I(X, Y_x) < I(X, Y_1 \cup Y_2 \cup Y_3 ... \cup Y_n) \tag{8.1}$$

If so, how could the theory handle this? Are there further restrictions we must introduce?

The data processing inequality

There is one type of indeterminacy challenge we can meet given the resources of the thesis. In section 6.6.1, I argued that indeterminacy of content between items in a causal chain - from most distal environmental item to most proximal state (i.e. the firing of the neurons immediately prior to the representation under investigation) - is not a problem

if we allow that the part of the chain *accessed* by the *system itself* is a random variable modelling a distal item, not a proximal item. In one sense, this provides an answer to what we might call the *distal indeterminacy problem*. In another sense, it is somewhat mysterious how the system itself manages to perform this feat of selecting from the distal end of the chain. This is especially true in light of the data processing inequality, raised by theorists as a problem for an informational theory of content (e.g. [Martínez, 2019]).

The data processing inequality, an information-theoretic concept, suggests that, in a Markov chain of random variables, $X \rightarrow Y \rightarrow Z$, Z cannot contain more information about X than Y. Information is only lost over channels but can never be gained. So, how does the system ensure that it gains information about X from Z, rather than information about Y?

The short answer is that perception is almost certainly not a Markov chain. A Markov chain is a stochastic process describing a sequence of events where the probability of an event later in the chain occurring is dependent only on the obtaining of an event earlier in the chain. This is almost certainly not the case for even basic perceptual systems, which have myriad feedforward *and* feedback connections from higher cortical regions (e.g. [Friston, 2003]). Stored information can be fed into the channel at various stages in order to allow a later stages, such as Z, to provide more information about more distal stages in the sequence, such as X, than proximal stages such as Y.

The brain, in essence, has its *own* way of solving the distal indeterminacy problem.

8.4.3 Translation into other models

Is maxMI compatible with other formal models, such as Kolmogorov complexity? In general, this is a question about how *sui generis* Shannon's information theory is, and whether other models can be used instead. Are those models better or worse in any sense?

8.4.4 Philosophical desiderata

How well does maxMI deal with various philosophical desiderata on representation? For example, can maxMI provide an account of what Jonathan Cohen calls "grain" [Cohen, 2004], or "format"? Does maxMI provide an account of the various *ways* in which content is represented? Is *encoding scheme* the same kind of thing as philosophers mean by format?

Generally, there are a plethora of possible philosophical questions we might ask about maxMI and its status as representational content in a rich philosophical sense.

8.5 Summary

The impilict theory of content in some regions of cognitive neuroscience is maxMI. The content of a representation is that item in the environment with which the representation has maximal mutual information, within the set delineated by the C-function of the subsystem housing the representation. The implicit theory captures the information about the environment available to the system itself, and provides a type of content which can feature in the theory proper of science.

I hope that this thesis can bring to the attention of some philosophers the sophisti-

cation of information theory, beyond the limits some have assumed it to have. I hope it highlights the complexity of the properly scientific use of content, and settles questions about the role that content plays in science, as well as the type of content and functions which can be used in proximal explanations.

I hope it can bring to the attention of cognitive scientists a way to systematise content ascription for particular projects, and to alert scientists to when content is used as a gloss. It may provide some helpful reflection on the discipline.

Content can be specified with a high degree of accuracy, with explanatory value, once we attend to the limitations of the system itself. The rest is noise.

Glossary

availability Information about an item is available if processing can be conducted over the values of the eRV used to model the item.

background theoretical framework a model of the properties and processes in that domain, which generalise across lower-level properties and processes which underlie that domain (e.g. physical interactions, chemical synthesis, etc.) using a set of concepts and principles which are independently well understood.

C-function Named after Robert Cummins, C-function is a non-etiological form of function. Defined by Cummins as follows:

> x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s's capacity to ψ just in case x is capable of ϕ-ing in s and A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to ϕ in s [Cummins, 1975, p762]

channel The physical medium of transmission of messages from sender to receiver. For example, the axon of a neuron or electromagmetic waves in the environment.

content An item external to a representation. Specifically, the source item of the receiver. In context, the item which is used in proximal explanations of a system's cognitive capacities. According to maxMI, the item with which the representation maximises .

encoding X encodes Y only iff there is some mathematical function f which 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 $x_{(1-n)}$.

entropy The average surprisal of a random variable, expressed as $H(X)$. Given by the formula

$$H(X) = -\sum_{i=1}^{n} p(x_i) log p(x_i)$$

eRV Abbreviation for **external random variable**: the random variable which models the content of a representation.

function The role performed, within a wider system, by a subsystem to enable the cognitive capacity which the subsystem serves.

informational teleosemantics A branch of teleosemantics emphasising an input condition on content. Defined by the use of information-relations between representations and contents in a theory of content.

iRV Abbreviation for **internal random variable**: the random variable which models the system-side representation.

item A term which is metaphysically neutral with respect to the properties of some aspect of reality. Loosely, a "thing" which exists.

maxMI The theory that the content of a representation is that item in the environment, modelled as an eRV, which maximises mutual information with a representation, modelled as an iRV, relative to other eRVs within the set delineated by the C-function of the subsystem housing the representation.

mutual information The amount of information in X which is about Y. Expressed as $I(X,Y)$ or $I(Y,X)$. Given by the formula

$$I(X,Y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} p(x_i, y_i) log \frac{p(x_i, y_i)}{p(x_i)p(y_i)}$$

naturalism Contrasted with pragmatic. Determined by scientific principles.

neural representation A representation for which the outcome values of its corresponding iRV are specified as states of single or multiple neurons, such as firing rate or voltage output.

noise A random channel value which interferes with the transmission of a signal from source to receiver.

pragmatic Determined by heuristic considerations. For example, communicability, arbitrary preference of explanatory project, perspicacity of relation to pre-theoretic interests.

proximal explanation An explanation of the system in terms of non-historical conditions. Aimed at answering how-questions. For example "how is face recognition performed?"

random variable "A random variable X is a function that maps each outcome x of an experiment (e.g. a coin flip) to a number $X(x)$, which is the outcome value of x." [Stone, 2015, p26]

realism The view that the formal language of a theory accurately describes some phenomena outside of that language. For example, the view that information theory accurately describes the interaction between the brain and the external world.

receiver The end of the communication channel, which processes decoded information. The downstream subsystem which decodes information from a representation.

representation A state, structure, or process mediating between an input and an output, and which has content.

source The item from which an eRV is modelled.

teleosemantics A branch of philosophy dealing with representational content. Defined by the use of functions in theories of content.

theory of content A theory which allows one to specify, once the relevant empirical facts are known, what the content of any given representation is.

theory proper That element of a scientific theory which is sufficient to explain the cognitive capacity under investigation.

ultimate explanation An explanation of the system in terms of historical conditions. Aimed at answering why-questions. For example "why does AM in IT represent $\langle shape \rangle$?"

W-function Named after Larry Wright, W-function is an etiological form of function. Defined by Wright as follows:

> The function of X is Z *means*
>
> (a) X is there because it does Z
>
> (b) Z is a consequence (or result) of X's being there. [Wright, 1973, p161]

Bibliography

[Almeida et al., 2020] Almeida, J., Freixo, A., Tábuas-Pereira, M., Herald, S. B., Valério, D., Schu, G., Duro, D., Cunha, G., Bukhari, Q., Duchaine, B., et al. (2020). Face-specific perceptual distortions reveal a view-and orientation-independent face template. *Current Biology*, 30(20):4071–4077.

[Armour-Garb et al., 2023] Armour-Garb, B., Stoljar, D., and Woodbridge, J. (2023). Deflationism About Truth. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition.

[Artiga, 2021] Artiga, M. (2021). Beyond black dots and nutritious things: A solution to the indeterminacy problem. *Mind & Language*, 36(3):471–490.

[Artiga et al., 2020] Artiga, M., Birch, J., and Martínez, M. (2020). The meaning of biological signals.

[Atick and Redlich, 1990] Atick, J. J. and Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural computation*, 2(3):308–320.

[Barsalou, 1999] Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

[Bedny, 2017] Bedny, M. (2017). Evidence from blindness for a cognitively pluripotent cortex. *Trends in cognitive sciences*, 21(9):637–648.

[Bergman, 2021] Bergman, K. (2021). Should the teleosemanticist be afraid of semantic indeterminacy? *Mind & Language*.

[Bergman, 2023] Bergman, K. (2023). Should the teleosemanticist be afraid of semantic indeterminacy? *Mind & Language*, 38(1):296–314.

[Bialek and van Steveninck, 2005] Bialek, W. and van Steveninck, R. R. (2005). Features and dimensions: Motion estimation in fly vision. *arXiv preprint q-bio/0505003*.

[Borst and Theunissen, 1999] Borst, A. and Theunissen, F. E. (1999). Information theory and neural coding. *Nature neuroscience*, 2(11):947–957.

[Bratman, 1993] Bratman, M. E. (1993). Shared intention. *Ethics*, 104(1):97–113.

[Brenner et al., 2000] Brenner, N., Bialek, W., and Van Steveninck, R. d. R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702.

[Bröhl et al., 2022] Bröhl, F., Keitel, A., and Kayser, C. (2022). Meg activity in visual and auditory cortices represents acoustic speech-related information during silent lip reading. *Eneuro*, 9(3).

[Burge, 2010] Burge, T. (2010). *Origins of objectivity*. Oxford University Press.

[Cao, 2012] Cao, R. (2012). A teleosemantic approach to information in the brain. *Biology & Philosophy*, 27:49–71.

[Cao and Warren, 2023] Cao, R. and Warren, J. (2023). Mental representation,"standing-in-for", and internal models. *Philosophical Psychology*, pages 1–18.

[Chang and Tsao, 2017] Chang, L. and Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028.

[Cohen, 2004] Cohen, J. (2004). Information and content. *The Blackwell guide to the philosophy of computing and information*, pages 213–227.

[Cohen, 2015] Cohen, J. (2015). Perceptual constancy.

[Cover and Thomas, 1999] Cover, T. M. and Thomas, J. (1999). *Elements of information theory*. John Wiley & Sons.

[Cummins,] Cummins, R. Neo-teleology. *Philosophy of Biology. An Anthology*, pages 164–174.

[Cummins, 2010] Cummins, R. (2010). *The world in the head*. Oxford University Press.

[Cummins, 1975] Cummins, R. E. (1975). Functional analysis. *Journal of Philosophy*, 72(November):741–64.

[Davidson, 2001] Davidson, D. (2001). *Essays on Actions and Events: Philosophical Essays Volume 1*. Clarendon Press.

[de Ruyter van Steveninck et al., 1997] de Ruyter van Steveninck, R. R., Lewen, G. D., Strong, S. P., Koberle, R., and Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805–1808.

[de Wit et al., 2016] de Wit, L., Alexander, D., Ekroll, V., and Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic bulletin & review*, 23:1415–1428.

[DiCarlo and Cox, 2007] DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341.

[DiCarlo et al., 2012] DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

[Dretske, 1994] Dretske, F. (1994). If you can't make one, you don't know how it works. *Midwest studies in philosophy*, 19:468–482.

[Dretske, 1981] Dretske, F. I. (1981). Knowledge and the flow of information.

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

[Egan, 2018] Egan, F. (2018). The nature and function of content in computational models. In *The Routledge handbook of the computational mind*, pages 247–258. Routledge.

[Egan, 2020] Egan, F. (2020). A deflationary account of mental representation. *What are mental representations*, pages 26–53.

[Elliffe, 2000] Elliffe, M. (2000). Performance measurement based on usable information.

[Fodor, 1987] Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*, volume 2. MIT press.

[Friedenberg et al., 2021] Friedenberg, J., Silverman, G., and Spivey, M. J. (2021). *Cognitive science: an introduction to the study of mind*. Sage Publications.

[Friston, 2003] Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352.

[Friston et al., 2006] Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of physiology-Paris*, 100(1-3):70–87.

[Gallistel, 2017] Gallistel, C. R. (2017). The coding question. *Trends in Cognitive Sciences*, 21(7):498–508.

[Gallistel, 2020] Gallistel, C. R. (2020). Where meanings arise and how: Building on shannon's foundations. *Mind & Language*, 35(3):390–401.

[Garson, 2019] Garson, J. (2019). There are no ahistorical theories of function. *Philosophy of Science*, 86(5):1146–1156.

[Godfrey-Smith, 1993] Godfrey-Smith, P. (1993). *Functions: Consensus without unity*. Pacific Philosophical Quarterly 74.

[Godfrey-Smith, 1998] Godfrey-Smith, P. (1998). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

[Green, 2017] Green, M. S. (2017). How much mentality is needed for meaning? In *The Routledge handbook of philosophy of animal minds*, pages 313–323. Routledge.

[Grider et al., 2019] Grider, M. H., Jessu, R., and Kabir, R. (2019). Physiology, action potential.

[Griffiths, 2006] Griffiths, P. E. (2006). Function, homology, and character individuation. *Philosophy of science*, 73(1):1–25.

[Hacohen, 2022] Hacohen, O. (2022). What are neural representations? a cummins functions approach. *Philosophy of Science*, 89(4):701–720.

[Hagoort and Indefrey, 2014] Hagoort, P. and Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual review of neuroscience*, 37:347–362.

[Hájek, 2007] Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, 156:563–585.

[Hinton, 1990] Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46(1-2):47–75.

[Hubel and Wiesel, 1959] Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574.

[Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106.

[Ince et al., 2017] Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., and Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human brain mapping*, 38(3):1541–1573.

[Ince et al., 2015] Ince, R. A., Van Rijsbergen, N. J., Thut, G., Rousselet, G. A., Gross, J., Panzeri, S., and Schyns, P. G. (2015). Tracing the flow of perceptual features in an algorithmic brain network. *Scientific reports*, 5(1):17681.

[Kirsh, 2006] Kirsh, D. (2006). *Implicit and Explicit Representation*. John Wiley Sons, Ltd.

[Kriegeskorte and Diedrichsen, 2019] Kriegeskorte, N. and Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual review of neuroscience*, 42:407–432.

[Laughlin, 1981] Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10):910–912.

[Lean, 2014] Lean, O. M. (2014). Getting the most out of shannon information. *Biology & Philosophy*, 29:395–413.

[Liang et al., 2019] Liang, J., Hou, L., Luan, Z., and Huang, W. (2019). Feature selection with conditional mutual information considering feature interaction. *Symmetry*, 11(7):858.

[Liu et al., 2022] Liu, M., Duan, Y., Ince, R. A., Chen, C., Garrod, O. G., Schyns, P. G., and Jack, R. E. (2022). Facial expressions elicit multiplexed perceptions of emotion categories and dimensions. *Current Biology*, 32(1):200–209.

[Lombardi et al., 2016] Lombardi, O., Holik, F., and Vanni, L. (2016). What is shannon information? *Synthese*, 193:1983–2012.

[Mann, 2018] Mann, S. F. (2018). Attribution of information in animal interaction. *Biological Theory*, 13(3):164–179.

[Mann, 2023] Mann, S. F. (2023). The relevance of communication theory for theories of representation. *Philosophy and the Mind Sciences*, 4.

[Marr, 2010] Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

[Martinez et al., 2005] Martinez, L. M., Wang, Q., Reid, R. C., Pillai, C., Alonso, J.-M., Sommer, F. T., and Hirsch, J. A. (2005). Receptive field structure varies with layer in the primary visual cortex. *Nature neuroscience*, 8(3):372–379.

[Martinez, 2013] Martinez, M. (2013). Teleosemantics and indeterminacy. *Dialectica*, 67.4:427–453.

[Martínez, 2019] Martínez, M. (2019). A mark of the mental: In defense of informational teleosemantics, by karen neander. *Notre Dame Philosophical Reviews*, 2019.

[Mattingly, 2021] Mattingly, J. (2021). *Information and Experimental Knowledge*. University of Chicago Press.

[Mayr, 1961] Mayr, E. (1961). Cause and effect in biology: kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, 134(3489):1501–1506.

[Mead and Tomarev, 2016] Mead, B. and Tomarev, S. (2016). Evaluating retinal ganglion cell loss and dysfunction. *Experimental eye research*, 151:96–106.

[Mesulam, 1990] Mesulam, M.-M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 28(5):597–613.

[Millikan, 1987] Millikan, R. G. (1987). *Language, thought, and other biological categories: New foundations for realism*. MIT press.

[Millikan, 1989a] Millikan, R. G. (1989a). An ambiguity in the notion "function". *Biology and Philosophy*, 4(2):172–176.

[Millikan, 1989b] Millikan, R. G. (1989b). Biosemantics. *The journal of philosophy*, 86(6):281–297.

[Millikan, 1990] Millikan, R. G. (1990). Compare and contrast dretske, fodor, and millikan on teleosemantics. *Philosophical Topics*, 18(2):151–161.

[Millikan, 2001] Millikan, R. G. (2001). What has natural information to do with intentional representation? 1. *Royal Institute of Philosophy Supplements*, 49:105–125.

[Millikan, 2013] Millikan, R. G. (2013). Natural information, intentional signs and animal communication. *Animal communication theory: Information and influence*, pages 3316–3603.

[Millikan, 2017] Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford University Press.

[Millikan, 2021] Millikan, R. G. (2021). Neuroscience and teleosemantics. *Synthese*, 199(1):2457–2465.

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

[Morgan and Piccinini, 2018] Morgan, A. and Piccinini, G. (2018). Towards a cognitive neuroscience of intentionality. *Minds and Machines*, 28:119–139.

[Nanay, 2014] Nanay, B. (2014). Teleosemantics without etiology. *Philosophy of Science*, 81(5):798–810.

[Neander, 2013] Neander, K. (2013). Toward an informational teleosemantics.

[Neander, 2017a] Neander, K. (2017a). Functional analysis and the species design. *Synthese*, 194(4):1147–1168.

[Neander, 2017b] Neander, K. (2017b). *A Mark of the Mental: In Defense of Informational Teleosemantics*. MIT Press.

[Orlandi, 2020] Orlandi, N. (2020). Representing as coordinating with absence. *What are Mental Representations?*, page 101.

[O'Toole, 2011] O'Toole, A. J. (2011). *Cognitive and computational approaches to face recognition*. The Oxford handbook of face perception.

[Paninski, 2002] Paninski, L. (2002). Convergence properties of some spike-triggered analysis techniques. *Advances in neural information processing systems*, 15.

[Papineau, 1984] Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, 51(4):550–572.

[Papineau, 2021] Papineau, D. (2021). Naturalism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

[Paulin and Hoffman, 2001] Paulin, M. G. and Hoffman, L. F. (2001). Optimal firing rate estimation. *Neural Networks*, 14(6-7):877–881.

[Pecher et al., 2003] Pecher, D., Zeelenberg, R., and Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. *Psychological science*, 14(2):119–124.

[Piccinini, 2022] Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics*, 16:846979.

[Pillow and Simoncelli, 2006] Pillow, J. W. and Simoncelli, E. P. (2006). Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis. *Journal of vision*, 6(4):9–9.

[Purves et al., 2001] Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C. K., LaMantia, A.-S., O McNamara, J., and Williams, S. M., editors (2001). *Neuroscience, 2nd edition*. Sinauer Associates.

[Quiroga et al., 2005] Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

[Ramsey, 2016] Ramsey, W. (2016). Untangling two questions about mental representation. *New Ideas in Psychology*, 40:3–12.

[Ramsey, 2007] Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

[Renner and Maurer, 2002] Renner, R. and Maurer, U. (2002). About the mutual (conditional) information. In *Proc. IEEE ISIT*, volume 364.

[Salmon, 1984] Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

[Schulte, 2018] Schulte, P. (2018). Perceiving the world outside: How to solve the distality problem for informational teleosemantics. *The Philosophical Quarterly*, 68(271):349–369.

[Schulte, 2023] Schulte, P. (2023). *Mental Content*. Cambridge University Press.

[Schulte and Neander, 2022] Schulte, P. and Neander, K. (2022). Teleological Theories of Mental Content. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition.

[Schwartz et al., 2006] Schwartz, O., Pillow, J. W., Rust, N. C., and Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of vision*, 6(4):13–13.

[Schyns et al., 2020] Schyns, P. G., Zhan, J., Jack, R. E., and Ince, R. A. (2020). Revealing the information contents of memory within the stimulus information representation framework. *Philosophical Transactions of the Royal Society B*, 375(1799):20190705.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27.3:379–423.

[Shannon, 1956] Shannon, C. (1956). The bandwagon. *IRE Transactions on Information Theory*, 2.1:3.

[Sharpee et al., 2004] Sharpee, T., Rust, N. C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural computation*, 16(2):223–250.

[Shea, 2007] Shea, N. (2007). Consumers need information: Supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research*, 75(2):404–435.

[Shea, 2018] Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

[Shea, 2020] Shea, N. (2020). Representation in cognitive science: replies. *Mind & Language*, 35(3):402–412.

[Souza et al., 2022] Souza, F., Premebida, C., and Araújo, R. (2022). High-order conditional mutual information maximization for dealing with high-order dependencies in feature selection. *Pattern Recognition*, 131:108895.

[Stone, 2015] Stone, J. V. (2015). Information theory: a tutorial introduction.

[Stone, 2018] Stone, J. V. (2018). *Principles of neural information theory*.

[Stryker, 1992] Stryker, M. P. (1992). Elements of visual perception. *Nature*, 360(6402):301–302.

[Tanaka, 1992] Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology*, 2(4):502–505.

[Tanaka, 1997] Tanaka, K. (1997). Mechanisms of visual object recognition: monkey and human studies. *Current opinion in neurobiology*, 7(4):523–529.

[Tovée, 2008] Tovée, M. (2008). An introduction to the visual system.

[Tsao and Livingstone, 2008] Tsao, D. Y. and Livingstone, M. S. (2008). Mechanisms of face perception. *Annu. Rev. Neurosci.*, 31:411–437.

[Van Rullen and Thorpe, 2001] Van Rullen, R. and Thorpe, S. J. (2001). Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural computation*, 13(6):1255–1283.

[Von Melchner et al., 2000] Von Melchner, L., Pallas, S. L., and Sur, M. (2000). Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature*, 404(6780):871–876.

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

[Wilson et al., 2007] Wilson, R. A., Barker, M. J., and Brigandt, I. (2007). When traditional essentialism fails: biological natural kinds. *Philosophical Topics*, 35(1/2):189–215.

[Wright, 1973] Wright, L. (1973). Functions. *The Philosophical Review*, 82(2):139–168.

[Yang, 2018] Yang, J. (2018). Information theoretic approaches in economics. *Journal of Economic Surveys*, 32(3):940–960.

[Zhan et al., 2019] Zhan, J., Ince, R. A., Van Rijsbergen, N., and Schyns, P. G. (2019). Dynamic construction of reduced representations in the brain for perceptual decision behavior. *Current Biology*, 29(2):319–326.