Communication and cooperation in evolutionary biology

A thesis submitted for the degree of Doctor of Philosophy
of The Australian National University.

This thesis is 89760 words long, including footnotes, captions and appendices.

This thesis is solely the work of its author. No part of it has been submitted for any degree, or is currently being submitted for any other degree. To the best of my knowledge, any help received in preparing this thesis, and all sources used, have been duly acknowledged.

Acknowledgements

For their patience, encouragement, acuity and trust, my supervisors deserve great credit. They are Justin Bruner, Kim Sterelny, Ron Planer and Matt Spike. Kim welcomed me as a visiting student, guided my research towards a judicious intersection of interests and ability, then took me on as a doctoral candidate. The project that would eventually become this thesis was begun in September 2013, at King's College London. My supervisors there, Nick Shea and David Papineau, shepherded me well during the first wobbly steps of research life. I have been lucky enough throughout the project to co-author several papers with local and far-flung collaborators. Warm thanks to Jessica Pfeifer, Ross Pain, Anton Killin, and Seán Roberts (along with the rest of the CHIELD crowd) for their fellowship, generosity and expertise. For conversations both live and epistolary, formal and informal, thanks are due to Ruth Millikan, Isobel Ronai, Madeleine Beekman, Karen Neander, Eva Jablonka, Jonathan Birch, Paul Griffiths, Arnaud Pocheville, Peter Godfrey-Smith, Cailin O'Connor, Dan Hutto, Rachael Brown, Carl Brusse, Heather Browning, Domi Dessaix, Bruno Ippedico, Jess Isserow, Dave Kalkman, Chris Lean, Brant Pridmore, Cameron Turner, Toby Solomon, Lachlan Walmsley, John Zerilli, Jeremy Strasser, James Willoughby, Tim Williamson (the ANU instantiation), Chris Bottomley, Shang Long Yeo, Justin D'Ambrosio, Daniel Stoljar, Frank Jackson, Manolo Martínez, Marc Artiga, Jessica Keiser, Susan Pennings, Emma Davies, Don Nordblom, Jesse Hambly, Hezki Symonds, Katie Steele, Colin Klein, Julia Henke Haas, Ben Haas Henke, T. Mark Ellison, and Siva Kalyan. For invaluable support throughout my time in the School of Philosophy, thank you Karen Smith, Steph Brinkworth, Sophie Napier, Narelle Quigley and Duška Kolundzija.

Most of this thesis, and all my published work to date, was written during regular Shut Up and Write sessions run by the Postgraduate and Research Students' Association – long may they continue. Much of the final manuscript was composed in the apartment of Aini Skirgård, Åsögatan, Stockholm – many thanks Aini! I gratefully acknowledge support provided by an Australian Government Research Training Program Scholarship and Australian Research Council Laureate Fellowship Grant FL130100141, without which the project would not have been possible.

Finally, to my partner, Hedvig Skirgård, and my family, Julie, Brian and Jenny Mann, my eternal love and unbridled gratitude. I dedicate this thesis to you.

This project was conceived, composed and completed on the lands of the Ngunnawal people.

Abstract

How can the concepts and results of communication theory aid evolutionary biology? This thesis argues for an explanatory framework, evolutionary communication theory, that interprets and illuminates scientific research into the phenomenon of biological signalling. By expanding the theory beyond the models and goals familiar to Claude Shannon and other engineers, real insight is gained into how strategic interplay between senders and receivers shapes signal form. Furthermore, interpreting artificial and natural signals in terms of sender-receiver teleosemantics demonstrates the explanatory role of relations borne between signals and world affairs. One of the major results of the thesis is a rejection of the orthodox distinction between Shannon and semantic information. While there are at least two useful distinctions to be drawn – between cues and signals, and between statistical and functional content – the terminological confusion that gave rise to the phrase 'Shannon information' should be put aside for good.

Chapter 1 outlines a way to capture the relationships between signals and other signal-like interactions using a multi-dimensional conceptual space called a hypercube. I argue that sender-receiver teleosemantics is uniquely well suited to capturing those aspects of communication theory that render it a viable mathematical framework for evolutionary biology.

Chapter 2 discusses an early attempt to apply communication theory in evolutionary biology. Haldane & Spurway's informational interpretation of the honeybee waggle dance has recently been criticised on mathematical grounds. These criticisms lend support to scepticism about the relevance of information for evolutionary biology. I argue that the criticisms are themselves mathematically erroneous, so one route to scepticism about information is

undercut. Chapter 3 explores a related line of scepticism. It is common in the philosophy of biology to treat the concepts and tools of communication theory as insufficient or irrelevant for analysing semantic content. I argue that the grounds of this supposition are based on misinterpretations of some features of communication theory.

In chapter 4 I reconstruct Millikan's teleosemantics in a causal-modelling setting, highlighting the explanatory role of semantic content. In chapter 5 I respond to objections to the teleosemantic account, including the claim that the theory renders explanations of success that appeal to semantic content circular. I also argue for an interpretation of important features of communication-theoretic models in terms of teleosemantics.

Chapter 6 explores another challenge to applying teleosemantics to biological signals. The theory places emphasis on cooperation between senders and receivers, but biological signals are often fraught with evolutionary conflict. I discuss recent formal work, and argue that prospects for teleosemantics are good. Finally, in chapter 7 I argue that an explanatory framework that draws on communication-theoretic concepts would be beneficial to evolutionary biology. I present case studies of communicative behaviour for which biologists offer explanations that are well interpreted through the principles of communications engineering.

Contents

Introduction

> What do we mean when we say that A communicates with B? We mean that the physical acts of A have induced a desired physical state in B.
>
> Cover and Thomas (2006, p. 183)

A bird sits on a telegraph wire and sings. What is the difference between its song and the electrical pulse within the wire? One is an acoustic wave produced by the bird's vocal organ, the other is an electric current produced by a transmitting station. One moves at the speed of sound and is relatively short-ranged, the other moves at the speed of light and can be relayed indefinitely. The form of the song is determined by various genetic, developmental, perhaps social, and probably accidental factors of the bird, while the form of the transmission is determined by a coding scheme invented as a result of decades of mathematical analysis – plus a dose of noise. That mathematical work originated in the early 20th century and matured immediately after the Second World War, with the publication of "A Mathematical Theory of Communication" by Claude Shannon (1948a). The definitions and theorems appearing in that text played a fundamental role in the technological advances of the late 20th century.

Telegraph lines existed before 1948, of course. But Shannon and his colleagues at Bell Labs developed a foundational approach to the engineering problem of communication. Even before its practical applications impacted public life, biologists became interested in using communication theory to analyse biological signals. The 1950s saw a flurry of symposia and publications aimed toward novel mathematical analyses of biological structures and systems (Attneave, 1954; Haldane and Spurway, 1954; MacKay and McCulloch, 1952; Quastler, 1953; Yockey, 1958). It seemed to many that Shannon was offering to quantify what had previously been only informally described. This excited a lot of people.

Sadly, no consensus emerged about how to apply communication theory in biology. Once the initial wave of excitement died, hope gave way to scepticism. Some biologists continued to quantify signalling behaviour using Shannon's tools. For the most part, however, the lure of precision wore off. It was simply not clear what would be achieved by translating data extracted from biological observations into the language of communication theory.

Contemporary philosophy of biology is still concerned with these issues. But its treatment of communication theory is inadequate. Sceptical arguments ignore its range of application. Its relationship to philosophical and biological accounts of communication is neglected. In this project I begin to redress these shortcomings.

The thesis is structured as follows.

Chapter 1 introduces the central model and sender-receiver teleosemantics. Both deal with signals, but evolutionary biology recognises several kinds of signal-like interactions. The chapter outlines a way to capture the relationships between them using a multi-dimensional conceptual space called a hypercube. I characterise signals, placing artificial and natural communication in a common form, and distinguish them from two signal-like interactions: cues and influence. I argue that sender-receiver teleosemantics is uniquely well suited to capturing those aspects of communication theory that render it a viable mathematical framework for evolutionary biology. I also introduce the properties and relations attributed to signals by sender-receiver teleosemantics. These will be crucial to later arguments.

In chapter 2, I explore an early attempt to model biological signalling in terms of communication theory. Haldane and Spurway (1954) use statistical data to measure the rate of information transmission in the honeybee waggle dance. Recently, philosophers have questioned their modelling choices and thus the relevance of their results. These criticisms lend

support to scepticism about the relevance of information for evolutionary biology. I argue that the criticisms are themselves mathematically erroneous, so one route to scepticism about information is undercut.

Chapter 3 aims to correct some core misconceptions about the application of communication theory in biology. It is common in the philosophy of biology to treat the concepts and tools of communication theory as insufficient for analysing semantic content. The strongest form of this claim states that communication theory is irrelevant to philosophical theories of content. Usually, such claims are supported in two ways: by citing Claude Shannon's warning that information ought to be distinguished from meaning, and by pointing out that statistical correlation is a broader category than semantic content. I argue that Shannon's warning ought to be understood in a way that does not prevent communication theory playing a role in a philosophical theory of semantic content. I further argue that communication theory has more concepts and tools at its disposal than just measures of statistical correlation.

Chapters 4 and 5 present the details of the teleosemantic theory of content. In chapter 4 I argue that relations can play an explanatory role in causal models, drawing on the interventionist explanatory framework. I then reconstruct Millikan's teleosemantics in a causal-modelling setting, highlighting the explanatory role of 'mapping relations'. Mapping relations are the relations that hold between signals and world affairs that teleosemantics identifies as the fundamental form of semantic content. In chapter 5 I respond to some objections to the teleosemantic account, including the claim that the theory renders explanations of success that appeal to semantic content circular. I also argue for an interpretation of some features of communication-theoretic models in terms of teleosemantics.

Chapter 6 considers extending teleosemantics to cover cases of signalling under divergent

interests. The theory places emphasis on cooperation between senders and receivers, but biological signals are often fraught with evolutionary conflict. I discuss recent formal work, especially results drawing on an aspect of communication theory called rate-distortion theory. I argue that prospects are good for extending teleosemantics.

Chapter 7 draws together many of the themes of the project in arguing for a novel label for a family of existing explanatory practices: **evolutionary communication theory**. The theory provides an explanatory framework for biological signals using communication-theoretic principles. The foregoing chapters cleared away the strongest forms of scepticism about the concept of information and its relevance of biology. The final chapter presents a positive story about the benefits of applying that concept, and other mathematical methods and tools, in evolutionary studies of signalling behaviour. Mainstream biology welcomes these principles, and it is time for philosophy to accept them too.

The remainder of this introduction covers important terminology.

Terminology and clarifications

Symbols. The following symbols denote features of signalling models: states $W = \{w_i\}_{i=1}^{l}$, signals $S = \{s_j\}_{j=1}^{m}$, and acts $A = \{a_k\}_{k=1}^{n}$. Matrices represent payoffs: $Q = q_{ik}$, where q_{ik} is the payoff for act a_k in state w_i. When sender and receiver have different payoffs, this is denoted by a subscript (Q_s, Q_r) or superscript (q^s_{ik}, q^r_{ik}). See also appendix B.

Discrete variables. All models in this thesis use discrete rather than continuous variables, with the following exceptions:

- The waggle dance model due to Haldane and Spurway (1954) (section 2.2)
- The costly signalling model due to Grafen (1990) (appendix D).

Informational equations and definitions, such as entropy, surprisal and statistical content, are presented in their discrete forms.

Case studies. Case studies appear alongside the main text in grey boxes. Although boxes often contain non-essential content, I use them as a way to separate scientific from philosophical material. All boxes contain content that can be considered essential to the discussion in the main text. The aim is to present the science free from philosophical bias, before discussing it in the main text with respect to my philosophical claims.

A case study

Information in these boxes should read as presenting scientific material free from philosophical perspective.

Published material. Certain previously published material appears in the text. This is taken from Mann (2018) and is indicated in each case.

Figures. All figures were created by the author using Paint Dot Net image editing software, Anaconda 3 with Matplotlib, or TikZ for LaTeX, with the following exceptions:

- Figure 1.5 on page 35 is a combination of figures taken from Churchland (1996, p. 24), Mitchell (2000, p. 263), Godfrey-Smith (2009, p. 64) and Dennett (2017, §7).
- Figure 2.1 on page 66 is taken from Chittka (2004, p. 898) and distributed under the terms of the Creative Commons Attribution License.
- Figure 4.4 on page 141 is taken from think-maths.co.uk, a collection of teaching resources designed for free distribution.
- Figure 5.6 on page 213 is taken from MacKay (2003, p. 9), freely available online.

Communication theory and information theory. Communication theory is the collection

of mathematical results and engineering methods that centre specifically around the fundamental problem of communication: reproducing a symbol string through a noisy channel (see appendix A). Information theory is a much wider class of mathematical concepts and results, applicable outside engineering to topics that have nothing to do with communication. Because my focus in this thesis is on biological signals, I often refer to communication theory rather than information theory; the terms are not synonymous.

Evolutionary biology. The term *evolutionary biology*, as appears in the title of the thesis and throughout, encompasses evolutionary approaches to explaining biological phenomena. This particularly includes accounts in behavioural ecology that seek to explain animal signalling by reference to evolutionary processes, as indicated by the title of the series *Behavioural Ecology: An Evolutionary Approach* (Krebs and Davies, 1978, 1984).

Chapter 1

Signalling in the abstract

1.1 Introduction

Signals have a fundamentally coordinative function. They mediate joint behaviour of two or more entities. As a result, they bear relations to world affairs relevant for the success of that behaviour. In biology, explanations of signalling behaviour refer to these relations. This chapter explores the idea that such explanations are special, and that what is special about them has something to do with the relations signals bear to the world. I argue that these ideas are best understood through the lens of **sender-receiver teleosemantics**, a theory of meaning.

My aim is to show that signalling is special relative to other functional behaviours. But biological signalling, which forms the majority of our case studies, is evolutionarily continuous with other behaviours. Signals lie on a many-dimensional continuum, and not all regions of the resultant space share their special explanatory status. For example, signals often evolve from other kinds of dyadic interaction, such as **cues** and **influence**. These three interactions differ along the *common interest* dimension of the continuum, just one of many dimensions that comprise the relevant space. If cues and influence are not apt for the same explanatory approach, where does the difference lie?

The space that characterises signals and related interactions is an example of a **hypercube**. Hypercubes are multi-dimensional continua used to elucidate relations between simi-

lar phenomena. The signalling hypercube is covered by a patchwork of distinct scientific **paradigms**, some of which overlap. Paradigms are scientific practices characterised by shared properties of the phenomena they study. Different scientists study different regions of the many-dimensional continuum that signals inhabit. For example, microbiologists studying quorum-sensing behaviour of bacteria are situated in a region of the hypercube characterised by typically (though not exclusively) high common interest, relatively high resource separation (see below), and somewhat arbitrary signal form. Behavioural ecologists studying animal mating behaviour are much further toward the 'conflict' end of the common interest dimension. The signals they consider may be more or less arbitrary than bacterial signals depending how we interpret that term (see section 1.5).

One consequence of this chapter will be that *explanations of biological signalling behaviour track the same features of signals required by communications engineering*. Engineers do not typically provide explanations, but they do offer justifications for design choices which, in a biological context, can inform scientific explanation. In other words, biology and engineering share an explanatory paradigm. My aim is to make plausible the idea that communication theory can do certain kinds of explanatory work when applied to biological signals. To argue for the claim, I will introduce several kinds of interaction and place each in its appropriate spot on the hypercube. By showing that communication theory shares a conceptual space with several different biological signalling paradigms, I will argue for the *prima facie* utility of communication-theoretic concepts in biology.

This is how the chapter will proceed. First we require a working definition of signal. I introduce one (section 1.2) before distinguishing signals from cues (1.3) and influence (1.4). Then, I present a series of conceptual spaces to highlight the evolutionary continuity between

these categories, as well as other kinds of interaction such as mechanisms (section 1.5). I argue that the paradigm of communication theory belongs in the same conceptual region as biological signalling. I then argue that sender-receiver teleosemantics is uniquely well-suited to demonstrating the links between historically different paradigms (section 1.6). Together, these considerations support the legitimacy of applying communication theory to describe and explain biological signals.

1.2 Signals

Signalling is a form of coordinated behaviour. An appropriate characterisation of signal must therefore rest on an appropriate characterisation of coordinated behaviour. The kind of behaviour we are interested in is not simply dispositional, not just causal effects an organism would have under various circumstances. Rather, it is functional behaviour, causal effects the organism was selected to have. Signals mediate functional behaviour, and are designed to do so.

For the sake of fixing our target, we may start with a minimal definition:

> A **signal** is an intermediary between two entities such that it is the proper function of one (the sender) to guide the proper function of the other (the receiver).

All cooperative biological communication counts as signalling under this definition.[1] Examples include the honey bee waggle dance, aposematic colouration from genuinely toxic prey to would-be predators, and scent trails left by eusocial insects. Furthermore, the definition picks out artificial signals too. Communication equipment designed by humans includes

[1] Maynard Smith and Harper (2003) and Scott-Phillips (2008) defend similar definitions in the context of animal communication.

transmitters and receivers. In being designed, these devices possess proper functions; in being designed to coordinate with each other, they satisfy our definition of signalling.

What is a **proper function**? It is a causal effect that an object or process was selected to have. Here *selection* encompasses any kind of selection process, including evolution by natural selection, learning through trial-and-error, and explicit design by rational agents. Because anything with a proper function falls into the relevant explanatory paradigm, signals are an ubiquitous and important feature of the *design stance* (Dennett, 1987). Treating objects in the world as possessing goals – treating them as **agents** as I will use the term – involves (among other things) evaluating their actions in terms of success and failure (Okasha, 2018, Introduction). Agents have goals and can fail to achieve them. For biological agents, the goal usually cited is optimisation of inclusive fitness (Lehmann and Rousset, 2014). Beyond biology, we can capture an agent's goals by citing its proper functions, what it was selected to do (Millikan, 1984, §1).

I explore these concepts in more depth in a moment, but first must deal with a more pressing issue. The minimal definition does not distinguish signals from what are intuitively 'mere mechanisms'. For example, the camshaft in a car satisfies our definition, but we do not want to call it a signal. Biologists too have mentioned the difference between, on the one hand, emitting a call to bring your offspring back to the nest, and on the other, physically carrying them back (Dawkins and Krebs, 1978, pp. 282–3). Counterexamples like this motivate augmenting our minimal definition. However, I do not want to make that move just yet. When I introduce the multidimensional space in section 1.5, it will have a prominent role for the *resource separation* between signals and the behaviours they guide. Paradigmatic signals are those for which all or most of the energy required to perform the act is provided by the

receiver itself, or at least at the receiver's end of the signalling channel. This goes some way towards capturing the intuition.[2]

In a biological context, the 'intermediaries' mentioned in the minimal definition take three main forms, which I will now introduce.

1.2.1 Three kinds of signal vehicle

Within our definition, there are three kinds of signal: behaviours, morphology and emissions.

First, **behaviours** can count as signals. For example, the waggle dance of the honey bee serves a joint function of sender and receiver by recruiting receivers to valuable food sources. Because of shared genetic interests, both parties gain when the receiver collects food efficiently. Compare behaviours that are not signals. Foraging behaviour is designed to help individuals find food. Foraging behaviour just as such is not a signal. Its function is not to support another organism's behaviour. Foraging behaviour might be observed by others, and they may change their own behaviour to improve foraging efficiency. Nevertheless, if the original foraging behaviour was not selected for this effect, it does not count as a signal.

Second, **morphological features** can be signals. Aposematic colouration in toxic frogs serves the function of warding off predators. Both frog and predator gain: the frog avoids being eaten and the predator avoids potentially deadly intoxication. Aposematic colours that are selected for this joint function are signals. Compare morphological features that are not

[2] Although I do not discuss intentionally produced signals, it is worth mentioning the importance placed on the so-called "causal inefficacy" of ostensive-inferential communication (Moore, 2017, p. 4). Moore interprets Sperber and Wilson (1986) as giving a Gricean account of communication, on which ostension – alerting receivers to the sender's communicative intention – is the mechanism by which signals prompt receiver action despite having negligible causal power. Similarly, Gupta and Sinha (2019, p. 2) characterise great ape gestures as "mechanically ineffective", highlighting the necessarily voluntary nature of the response and suggests socially regulated (i.e. broadly cooperative) interaction. Analogously, my claim here is that paradigmatic biological signals are relatively low-energy when (and because) sender and receiver are coadapted.

signals. The length of an animal's leg may serve a function related to its wider skeletal structure and typical motion. It is not thereby a signal, because it is not designed to engender some beneficial effect in another organism. Observable morphology may fortuitously benefit other organisms. A predator might assess which prey are more likely to be harder to chase by comparing the length of their legs. Nonetheless, if leg length was not selected to aid this assessment it is not a signal. Longer legs may have the function of helping the animal run faster, and predators may be the selection pressure underpinning this design. But a feature selected for effect X is not necessarily a feature selected to inform another organism that X.

Finally, physical **emissions** can count as signals. For example, chemical trails laid by foraging ants lead nestmates to food. Emissions differ from behaviour and morphology in that they can be physically discontiguous from their sender. They contrast with non-signalling emissions in predictable ways. Waste products, whether solid, liquid or gaseous, are produced by all organisms. Such emissions are not signals because they do not have a function to cause a beneficial impact to some receiver. As before, waste products may be opportunistically used by other organisms to their benefit. But unless their production is in part selected for this downstream benefit, waste products are not signals.

The three types of signal can be mapped onto two causal models (figure 1.1). Behavioural and morphological signals are represented by a model with two nodes corresponding to sender and receiver. There is no distinct causal variable representing the signal itself. Behaviour is a property of a sender, and morphology a configuration of it, rather than either of them being causally downstream. In contrast, emitted signals are physically distinct from senders and so can be represented as a third node.

Despite this difference, we will for simplicity use the common three-node format in de-

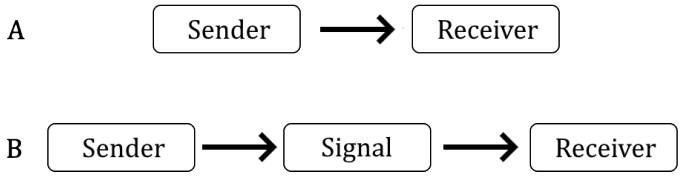

Figure 1.1: Signals represented as causal models. **A** Two-node model. **B** Three-node model.

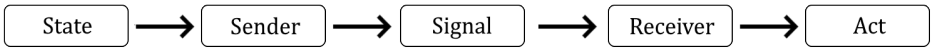

Figure 1.2: The causal model at the heart of the sender-receiver framework (see also appendix B).

scribing signals. The three-node model is a component of the **sender-receiver framework**, a family of models designed to aid theorising about signalling behaviour in nature. To our three nodes it adds a *state*, observed by the sender, and an *act*, performed by the receiver (figure 1.2). Senders and receivers gain a benefit or suffer a penalty when certain acts are performed concurrent with certain states obtaining. Intuitively, if benefits and penalties for sender and receiver coincide, both agents should be incentivised to coordinate their behaviour with the use of the causal intermediary: the signal.[3] The sender-receiver framework allows us to capture many important examples of signalling in nature. It highlights the relationship between these cases and the **central model** of communication theory, introduced in the following case study.

[3]Although the sender-receiver framework is usually associated with game theoretic models, it draws on situations that can be represented as causal models. For our purposes there is no difference between signalling games, sender-receiver games and sender-receiver models.

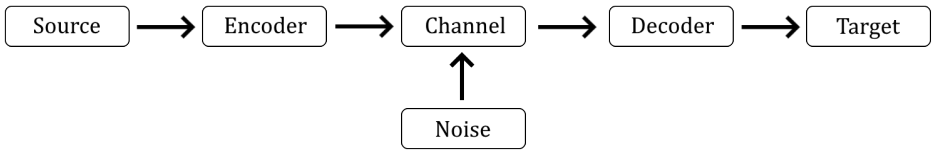

Figure 1.3: The central model of communication theory. See also appendix A.

The central model of communication theory

During the Second World War, an engineer named Claude Shannon devised an abstract mathematical framework for representing communication. Shannon (1948b) conceptualised the fundamental goal of communication as reconstruction of a symbol string. He presented a formal model of that goal and a way to measure success, setting the stage for an explosion of digital technology in the second half of the twentieth century. Throughout this thesis, the term 'central model' will refer to this model (figure 1.3).

Shannon's model contains six nodes (figure 1.3). At the **source**, one from a set of possible symbol strings is selected. The **encoder** translates the selected string into a codestring and transmits it across the **channel**, at which point it is subject to transformation due to **noise**. The **decoder** receives the (possibly noise-affected) codestring and must attempt to reproduce the originally selected string at the **target**.

It is assumed the decoder has access to the statistics with which the string is generated, so that even without the signal it could make an educated guess. But the signal provides a causal link with the actual string, improving the receiver's chance of success. Not just any connection will do: receivers need to know *how* to condition their choice on the signal. They do this by sharing a **code** with the sender.

Codes are rules for converting source strings into signals and back again. Signals in the central model are called **codewords**. Symbols in the source lexicon are translated into the code lexicon, which usually consists of 1s and 0s. These are then interpreted by the receiver into the original symbols.

In realistic channels, noise may interfere with the signal. Some of the symbols of the codeword are flipped. Anticipating this, the code should enable robustness, allowing the receiver to make the best possible decision given a noisy signal containing erroneous symbols. Appendix A describes concepts and results that this model enables.

The central model can be thought of as *a special case of an extension to* the sender-receiver framework. It is an extension because it adds a node representing **noise**. It is a special case because sender and receiver share a goal – they have perfectly aligned interests – and this goal is in particular reconstructing a symbol string. The string is generated randomly at the state node, and the receiver must reconstruct it.

1.2.2 Proper function

This thesis concerns two kinds of signals – biological and artifical – and the analogies between them. The key feature they share is being designed, in different though well-understood senses. Biological signals are a consequence of natural selection, learning, and sometimes (though rarely) other kinds of selection process such as cultural evolution. Artificial signals are designed by conscious, rational human engineers.[4]

[4]One might reasonably wonder which design process(es) give rise to human *natural language*: biological, ontogenetic, cultural, rational, some combination of the four, or others not listed? I do not pursue this question here. I suspect that – in part due to similarities in the selection processes underlying them – biological and artificial signals are more like each other than either is like language.

The design processes that produce signals, though importantly different in many ways, are all **selection processes**. Fundamentally, selection processes generate **proper functions**. Signals are a special case of proper function, a case in which behaviours of two entities correspond as a result of selection acting jointly on them.

Proper function is a term of art. It marks a large class of items in the natural world, and many in familiar human contexts. What is important for our purposes is that the same kind of pattern, of joint proper function, can be found in all our examples of signalling. Joint proper function – or just **joint function** for short – grounds the unique explanatory role of signals. Joint functions are functions belonging to interactions between two devices. Mutually beneficial interactions contribute to reproduction of types that take part in the interaction. This in turn contributes to the recurrence of that type of interaction. Signals explain their own existence by taking a form that mediates beneficial behaviour of sender and receiver.

To be sure, at its broadest, this explanatory pattern is common to any functional behaviour. The form of a functional behaviour explains persistence of the entity with that function – that is just what the concept of proper function captures. Joint functions concern two entities with coordinated behaviour, where the form of each explains persistence of the interaction.

In the remainder of this section we will focus on situations of **perfect common interest**, where joint functions are not threatened by individual incentives to exploit. Perhaps such scenarios rarely or never exist in nature; nonetheless, understanding the hypothetical cooperative case is crucial to understanding real phenomena. Divergent interests, which are admittedly crucial in many explanations of signalling behaviour, are considered in chapters 6 and 7.

Signals, then, are special because they *mediate* joint functions. This means that their

paradigmatic form is informational rather than mechanical. This in turn warrants attribution of a philosophically vexed feature: mapping relations between signals and world affairs.

1.2.3 Mapping relations

Because signals cannot physically force receivers to perform an action, there must be a reason why receivers choose the actions they do in response to signals. **Teleosemantics** says that this reason can be found by positing a relation between the signal and the circumstances in which the receiver acts. Call these relevant circumstances the **signified**. In engineering and science, the relationship between signal and signified is typically described by saying one *carries information about* the other. One of the contentions of this thesis is that this locution is best understood through teleosemantics. In particular, it should be understood in terms of sender-receiver teleosemantics as introduced by Ruth Millikan (1984) and further defended by Millikan (1989, 2004b, 2017), Artiga (2014b, 2016a), Martínez (2019) and others. Hereafter the term *teleosemantics* shall refer to this theory unless otherwise specified.

According to teleosemantics, receivers are able to perform their proper functions in the right way only when signals bear a particular relation to signifieds. Here the 'right way' to perform a function is, roughly, in the same way that a majority of the receiver's ancestors performed it when they contributed to success of their lineage through selection. When historically selected receivers were selected in part by responding to signals that bore relations to success-relevant world states – and those relations help *explain* how those receivers got selected – those relations are called **mapping relations**. A signal observed by a present receiver must bear that same relation, if it is to cause success in the right way. This being-supposed-to-bear-a-relation is what teleosemantics identifies as semantic content. A signal's

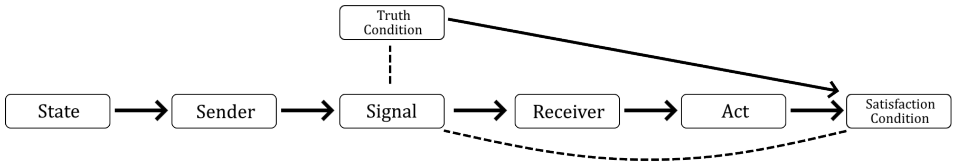

Figure 1.4: The basic teleosemantic model. Dashed lines represent mapping relations. The signal bears a descriptive relation to its truth condition and a directive relation to its satisfaction condition. Adapted from Millikan (2004b, fig. 6.3 p. 78).

mapping rule picks out the world state that must obtain in order for the signal to bear the relevant mapping relation to it. This brief exposition is expanded in great detail, along with an account of how relations in general can be explanatory, in chapter 4 and 5.

There are two kinds of fundamental mapping relation, directive and descriptive (figure 1.4). Directive mapping relations hold between a signal and what it is supposed to cause the receiver to bring about – its satisfaction condition. Descriptive mapping relations hold between a signal and what must be the case in order for the receiver's behaviour, conditioned on the signal, to be successful – its truth condition. For example, the directive aspect of the honeybee waggle dance maps onto the intended future behaviour of receiver bees in flying toward the indicated food source. The descriptive aspect is the location at which food would have to be present in order for receiver bees' subsequent behaviour to be successful.

Paradigmatic explanations of biological signalling behaviour track mapping rules. For example, a very simple mapping rule is the threshold effect of quorum sensing in communal bacteria like *Pseudomonas aeruginosa*. When colonising a new area like soil or a mammalian lung, *P. aeruginosa* coordinates behaviour to achieve more effective virulence. It benefits all individuals to act in accordance with the majority. Although quorum sensing is of most interest to philosophers as a collective action problem, it also provides an example

of a simple mapping rule. Sufficient density of the signalling molecule in the environment indicates that the requisite number of individuals are present. Equally, it directs individuals to begin whatever behaviour is being coordinated. The system may be portrayed as follows:

Concentration above threshold $\rightarrow_{\text{descriptive}}$ requisite number of individuals present

Concentration above threshold $\rightarrow_{\text{directive}}$ begin coordinated behaviour

Perhaps even simpler are the mapping rules of undifferentiated alarm calls:

Call here, now $\rightarrow_{\text{descriptive}}$ danger here, now

Call here, now $\rightarrow_{\text{directive}}$ take evasive action now

Mapping rules are also required for communications engineering to be possible. Communications engineers specify mapping rules in the form of codes. Encoding and decoding procedures induce mapping relations between codewords and world affairs. In the central model, as in most expositions of communication theory, the world affairs to which codewords map are strings of symbols. Nonetheless, mapping rules and the relations they determine are present for the same reason as the biological case: the signalling strategy was selected to aid coordination. In section 5.5 we will see that mapping rules are not identical with encoding and decoding procedures. In the special case of the central model, however, they coincide.

Many biological signals, and most of those we will deal with in this thesis, are **primitive**: directive and descriptive aspects are tightly coupled and both contribute equally to explanations of behaviour. The waggle dance is primitive because its indication of the presence of food is inextricably linked with its injunction to fly to that location. Likewise, signals in the

central model are primitive. Codewords indicate source symbols and simultaneously direct receivers to reproduce those symbols.

In sum, biology and communications engineering employ a common concept of signal grounded in a common concept of function. In both cases mapping rules illuminate the special explanatory role of signals. We therefore have *prima facie* reason to consider the application of communication theory in biology. Now, let us turn to a wider cast of characters. Understanding signals and their role in explanation requires understanding cues and influence too.

1.3 Cues

Behavioural ecology has a category for would-be signals whose sender is not coadapted with its receiver. They are called **cues**. Cues are sensory inputs on which receivers condition their behaviour, but which were not produced for the purpose of being used. They may be by-products of purposeful behaviour, as when mosquitos follow clouds of carbon dioxide emitted by mammalian respiration. Or they may be environmental affordances, as when plants track the position of the sun over the course of a day.

At several points in the discussion it will be useful to have a term covering both signal and cue. Let us designate the term **sign** for this purpose.

Whether a given sign counts as a signal or a cue is relative to the coevolutionary status of sender and receiver. An event that is a signal relative to a given receiver may be a cue relative to a different receiver. For example, the territorial call of the collared sparrowhawk is a signal to other sparrowhawks, because both the call and its subsequent aversion response

are coadapted. But the same call heard by a brown thornbill is a cue, because the response of the thornbill – increased vigilance for its sparrowhawk predator – did not contribute to the evolutionary success of the call (Schneider and Griesser, 2013, p. 48).

Cues are sometimes called natural signs (Millikan, 2013a). The term 'natural' is apt because any actor in a game theory setting whose behaviour is strictly independent from all other agents can reasonably be called 'Nature'. These agents comprise a fixed background against which focal agents operate. In sender-receiver models, for example, Nature is a player whose choices are random and are not affected by the choices of the other players. By definition, Nature's behaviour cannot be codesigned with that of other players.[5] Players that condition their own behaviour on Nature's are reading cues.

1.3.1 The signal/cue distinction is explanatorily relevant

Because a signal is codesigned, aspects of its form can be explained by reference to its effects. The same cannot be said of cues, because their form is not purposefully shaped for the receiver's benefit. This is a consequence of signals having proper functions.[6] We defined signals in terms of the functions of other entities (both senders *and* receivers), but they themselves have functions too. Like all functional behaviours, their form can be explained by reference to the specific effects their form is designed to have. Cues do not share this curious property, and biologists understand the relevance of the distinction. In response to commentary from Sasha Dall and other workers in ecological theory, Étienne Danchin and

[5]The term *codesign* is to be read as 'co-design', not 'code-sign'. Confusingly, both 'code' and 'sign' have important meanings in this thesis. The term 'code-sign' is meaningless and shall not be used, despite LaTeX's unfortunate tendency to hyphenate the term that way when split across lines.

[6]A complication: signals used as cues by eavesdroppers may possess features that promote secrecy. Then their status as cues is indeed relevant for their design. Thanks to Ron Planer for pointing this out.

colleagues agree that "by separating cues and signals, we can better understand both in order to subsequently synthesize them" (Dall et al., 2005, p. 355).

Despite this difference, cues bear a kind of correspondence to world affairs too. Scientists cite these in explanations of cue-reading behaviour. For example, mosquitos are attracted to mammals by sensing the carbon dioxide they emit due to respiration. There exists a tight spatiotemporal correspondence between carbon dioxide and mammal presence, and this explains the fact that mosquitos successfully, non-accidentally locate their target.

Cues in general bear **correspondence relations** to signifieds.[7] These relations are familiar from philosophical accounts of natural information (Stegmann, 2015) or natural meaning (Grice, 1957). Cues bear correspondence relations as a consequence of being maintained or produced by stable or recurrent processes. For example, Polaris serves as a cue indicating the direction of North because its position in relation to the tilt of the Earth's axis stays the same over long periods of time. Smoke serves as a cue indicating the presence of fire because it is reliably produced by the process of oxidation, which consists of burning fuel.[8]

1.3.2 Correspondence relations are neither correlations nor mappings

Correspondence relations must be distinguished both from correlations and from mapping relations. First, correspondence relations are not correlations. To spell this out a bit more, I am introducing 'correspondence relation' as a theoretical term referring to the relation borne

[7] My use of the term 'correspondence relation' is meant to capture what Millikan (2017) talks about as 'natural information'. This is perhaps a surprising use of the word 'correspondence', especially given its connotations in the history of philosophy of mind and language. I simply need a term to distinguish the kind of relation borne by cues from that borne by signals, and I chose 'correspondence' because to my ear it has a less functional sound than 'mapping'. It is perhaps an unsatisfactory choice, but I haven't found a better one yet.

[8] Millikan (2013a) gives the name "root signs" to cues considered with respect to their correspondence relations; Fresco et al. (2018) use the term "datum".

by a natural sign toward its signified, as described in the theory given by Millikan (2017). I treat 'correlation' in sense (c) of the Oxford English Dictionary (2020): "an interdependence of two or more variable quantities such that a change in the value of one is associated with a change in the value or the expectation of the others". In particular, there is no specific mathematical definition of 'correlation'; it denotes *any* statistical relationship between two variables. The clearest distinction between correspondence relations and correlations is that the former are borne between tokens – individual signs and their signifieds – while the latter is any measure of a relation between *types* (classes of tokens considered as statistical variables). Further differences will become apparent as we proceed.[9]

Before moving on to the distinction between correspondence relations and mapping relations, it is worth flagging two common misconceptions: first, the idea that cues ought to be distinguished from signals by reference to correlation; second, that correlations are necessarily *spatiotemporal* relations. First, it is a mistake to distinguish cues from signals by defining the former in terms of correlation. Correlations can be borne between signal and signified too, at the level of types. Later (section 3.4) I will argue that a popular measure of correlation, *mutual information*, has a different interpretation when applied to the signal-signified relationship than when applied to the cue-signified relationship. Second, it is wrong to treat correlations as having something to do with spatiotemporal relations. For example,

[9]The relationship between signs and their signifieds is not always as clear cut as the phrase 'token-token' suggests. An examiner pointed out that mosquitos may use clouds of carbon dioxide as a sign of a group of nearby mammals. There is not obviously a token-token relationship between any given molecule of carbon dioxide and any given mammal (or if there is, such a relation need not be mentioned in an explanation of the mosquito's behaviour in moving towards the group). I suspect this issue can be resolved by changing the level of grain at which cue and signified are specified: initially, the entire cloud is a single token; as the mosquito approaches, individual puffs emitted by a single mammal allow the mosquito to lock on and bite. In any case, I treat this as a problem of model choice (i.e. a problem for practitioners), rather than a problem for the salience of the modelling framework I employ throughout this thesis (i.e. a problem for theorists).

it is sometimes implicitly assumed that type *A* correlates with type *B* just in case *A* occurs at the same time (and perhaps also location) as *B*. However, nothing in the definition of correlation requires that the relationship between statistical variables be due to their time or place of occurrence.

Now to distinguish correspondence relations from mapping relations. The former are (by definition) borne as a consequence of natural processes other than joint selection, while the latter are (by definition) a consequence of joint selection. Further, because signals have functions which they can fail to perform, there is a reasonable sense in which mapping relations can fail to be borne. Because mapping relations are determined by mapping rules, they can fail. Since correspondence relations are not determined by rules but by processes, they cannot 'fail to obtain' in the same sense. These definitions, with slightly altered terminology, are intended to be faithful to those given by Millikan (2017, Part II) (for correspondence relations, there called "natural information") and Millikan (1984, Part II) (for signals, there called "intentional icons").

In short, while there are cues and non-cues there are no 'faulty cues', whereas there are indeed faulty signals. This is the position I take in the thesis, though there is disagreement in the wider literature. One tradition considers anything an organism treats as a cue to be a cue. On this view, there can be faulty cues: events a receiver treats as a cue but which do not bear the relevant correspondence relation. For example, carbon dioxide produced by an inorganic chemical process that a mosquito treats as a sign of mammalian presence would be a false cue. The approach I favour reserves the notion of falsity for signals only. Purported cues that fail to bear the correspondence relation are simply not cues – or at least not cues of the signified in question. Carbon dioxide is not a cue indicating the presence of mammals

when it does not bear the right relation to a mammal – namely, the relation of having just now been produced by one.[10]

The view that there cannot be false cues flows neatly from the fact that signals have functions and cues do not. There is no sense in which smoke produced by a smoke machine is supposed to map to fire. But there *is* a reason why a *signal* intended to prompt a fire-relevant response is supposed to map to a fire, even when there is no fire. The signal has a proper function, and to perform that function non-accidentally it must map onto a fire. If there is no fire, the signal fails to map. Cues are defined by correspondences they in fact bear, as opposed to those they are supposed to bear. To be sure, there are tokens produced by nature that animals misuse or mistake for cues. But they are not cues simply in virtue of mistakenly being used. For the remainder of the thesis I follow this tradition.

1.3.3 The signal/cue distinction is explananda-relative

In models of agential interaction, the signal/cue distinction is a matter of whose design is represented. Considered as part of a codesigned system, an intermediary is a signal. If the design of the sender is ignored or unspecified, it is only a cue. However, if the sender is treated as a feature of the environment, the signal/cue distinction disappears altogether. Models of decision-making, in which a receiver conditions its behaviour on sensory input, sometimes do not specify whether the sender was designed to produce that input. These models ignore the distinction between signal and cue.

[10]There is a residual question about certain forms of mimicry: some weeds, for example, mimic the visual appearance of domesticated crops and thereby successfully avoid destruction by humans. Plant appearance is a cue to the human (or can be assumed so for the sake of argument); in mimicking this, the weed could be said to be a source of misinformation. I suspect this kind of case can be resolved by treating the weed's appearance as a kind of false signal, even though the informational vehicle it is mimicking is a cue rather than a signal. Issues of this kind are discussed further in chapter 6.

If the signal/cue distinction is relative to what is being explained, it follows that the distinction between mapping relations and correspondence relations is relative too. When we ignore codesign, we ignore the mapping rules that pick out the states that the signals are supposed to bear relations to. Thus there are no mapping relations. However, in being converted into a cue, the former signal may still bear correspondence relations. *Treating the sender as part of nature removes its role in determining mapping rules, but gives it a new role as mediator of correspondence relations.*

Another way to describe the same perspectival shift is to say that sender behaviour is relegated to the physical stance. As a node in a causal model, it is interpreted not as an agent with a proper function but as part of Nature. Non-accidentally true signals become cues. False signals, along with those accidentally true, become non-cues, because cues are defined in terms of the natural processes that maintain their correspondence with signifieds. For former signals, sender behaviour previously considered proper functioning now becomes a natural process. We might say that ascending to the design stance is a case of figuring out how to distinguish false signals from non-signals. It is by appeal to the sender's function that we achieve this.

All I am trying to show is that there is a consistent way to treat signals as cues by ignoring or holding fixed the sender. Organisms can be modelled as agents or as part of the environment. Allowing for this switch of perspective is important, since it highlights the need for a consistent account of the emergence of signals from cues.[11]

[11] Does this entail that the definitions of signal and cue are merely pragmatic or instrumental? No: senders of signals really do have coadapted proper functions, even though we may model them as non-functional physical processes. That we can do this raises questions about proper functions, however – whether they themselves are real or merely instrumental. I treat them as real, and leave as an open question how it is possible for scientists to consistently and usefully switch between the physical and design stances.

Some explananda simply do not require consideration of the coevolutionary status of signs. From the perspective of these explananda, there is no signal/cue distinction. A scientist may be interested in the physiological mechanisms by which sensory input is transformed into behavioural response. They might even be interested in whether such behaviour is good for the receiver. Establishing the latter requires adverting to some correspondence between sign and world. For example, the sign might indicate that the sender is able to fight. The receiver's fleeing was adaptive for it on this occasion. But this alone does not tell whether the sign is a cue or a signal; it does not tell us whether the correspondence is of one kind or another. The difference is immaterial for these explananda. On the other hand, if we are interested in why receiver behaviour is reinforced, or why the form of the signal is what it is, we must refer to codesign.

Mathematical results that apply to cues properly belong to decision theory. That framework is the study of individual agents and how their behaviour depends on utilities, background knowledge and cues available in their environment (Jeffrey, 1990; Savage, 1954). Communication theory is closely related to both decision theory and game theory. In effect, the central model is a game in which both players have common interest and must collaborate on the task of reconstructing symbol strings (more on this in section 3.2). Game theory describes multiple agents whose behaviours are determined by natural or rational selection, and whose actions affect other agents (Maynard Smith, 1982; von Neumann and Morgenstern, 1944). Decision theory deals with one agent at a time, navigating an environment in which cues and perhaps signals are available to guide behaviour. Several results in communication theory do not require codesign, and so apply to cues as well. Signal detection theory is a collection of tools at the intersection between the two: it properly belongs to decision

theory, but its name betrays historical ties to communication theory (Wiley, 2013b).

All that being said, the distinction between signals and cues is continuous. Many signals start their lives as cues, slowly attaining proper functions as a result of coadaptation between sender and receiver. Rather than searching for a threshold at which a cue suddenly gains a proper function, we would do better to place familiar examples on a continuum. The explanatory distinction between signals and cues is real despite the vague boundary between them.

1.3.4 Ritualisation turns cues into signals

Cue-reading behaviour is a probable evolutionary origin of many signals. In these cases, receivers initially gained a benefit by conditioning responses on some recognisable feature of senders. This feature subsequently came under evolutionary control, when selection favoured those senders with more elaborate versions of the trait. The name for this evolutionary route is **ritualisation**. For example, warning colouration of toxic frogs may have begun as an accidental correlation between morphology and toxicity (Lehmann et al., 2014). It was in their predator's interest to heed such colours. They either learned or were selected to avoid those would-be prey. This is cue-reading. Subsequently, selection favoured elaboration of recognisable colours, and frog morphology changed to make recognition more probable. This is signalling.

There are two connotations of the term 'ritualisation'. The first is the evolutionary story just given. The second is the repetitiveness or stereotypy commonly observed in signals that followed this evolutionary path. When a receiver possess a conditioned response, and that response is in the sender's interest to elicit, the latter may redeploy the signal many times in

quick succession. Much theoretical work surrounds the specific form of signals of this kind (Maynard Smith and Harper, 2003, §5). Wiley (1983, §5.2.2) notes that ritualisation in this sense is reminiscent of what in communication theory is called **redundancy**, improving the prospects of an apt response. This provocative link is explored further in section 5.5.

To conclude, the distinction between signals and cues is robustly defined and explanatorily salient. There remains, however, a continuum between the two categories. The hypercube method (section 1.5) is a good way to capture their relationship. Before then, we need to explore a third category of functional interaction: influence.

1.4 Influence

Receivers condition their behaviour on cues, despite not being coevolved with the sender of the cue. Similarly, **influence** is not a coadapted behaviour, but is performed by a sender in order to control receiver behaviour. Behavioural ecology has long recognised these three kinds of joint behaviour. Calling signals *honesty*, cues *mind-reading*, and influence *manipulation*, Guilford and Dawkins give the following account:

> Honesty, manipulation and mind-reading, then, can all be seen as strategic components of signal design or, in the case of mind-reading, as a strategic component of non-signal design. Only the receiver benefits in mind-reading, only the emitter in manipulation but both benefit in honesty.
>
> Guilford and Dawkins (1991, p. 10)

Like many behavioural ecologists, Guilford and Dawkins consider both honesty and manipulation to fall within a category which they label "signal". What is the rationale behind this categorisation? Signals and influence have in common that they are produced because of

	evolved owing to the effect on the sender	benefits the receiver to respond
signal	+	+
cue	–	+
influence	+	–

Table 1.1: Distinguishing strategic interactions. Adapted from Diggle et al. (2007b, Table 2, p.1242) (they use the term "coercion" for what I call influence). The original caption reads: "Types of communication are distinguished depending upon their fitness consequences to the sender and the responder. (Consequences are either beneficial (+) or costly (–).)"

their functional effects. This is why several prominent accounts of biological communication, especially in behavioural ecology, consider both cooperative signalling and manipulative influence as signals. This categorisation also owes something to the fact that many common examples of animal interaction are broadly manipulative. Because 'signal' is the generic term applied to all such behaviours, it seems natural to draw a distinction between cues on the one hand and both cooperative and manipulative interactions on the other.

Whatever the benefits of a category that covers signals and influence but not cues, I prefer to distinguish all three. Consider table 1.1, taken from Diggle et al. (2007b, Table 2, p.1242). Each interaction is typed by the benefit or detriment it brings to participants. This is a useful way of distinguishing interactions because it follows the contours of explanation available for communicative behaviour. We should expect different interactions to possess different physical features and to follow different trajectories over evolutionary time. For example, truly cooperative signals that are not threatened by defection should be optimised for low energy expenditure. The only noise such signals need to overcome is that present in the environment. In contrast, influence will have features designed to overcome the receiver's reluctance to respond. In these cases, the receiver is a further source of disruption from the

perspective of the sender. Certain features of the receiver act as a barrier to the sender getting what it wants. The sender must optimise the signal to bring about the desired response. This includes expending energetic resources.

This distinction, between costs paid to overcome noise and costs paid to overcome receiver reluctance, is reflected in the literature as the distinction between **efficacy design** and **strategic design** (Guilford and Dawkins, 1991) (Maynard Smith and Harper, 2003, §5). Both signals and influence display efficacy design, because both need to reach receivers. On the other hand, strategically designed signals are those that correspond to some world affair, and elicit a response because of this correspondence. Influential behaviours do not have the same feature. Their strategic design consists of features designed to elicit a response in the receiver without necessarily corresponding to anything beneficial for the receiver.[12] Finally, cues are not designed to bring about a receiver response, so their physical features cannot be explained by reference to this kind of design. Indeed, many cues emitted by animals face opposing selection pressure. Exploitation is harmful, and they would do better to be hidden. Camouflage techniques and certain kinds of mimicry are canonical examples.

In sum, our three kinds of interaction are distinguished by the selection pressures acting on them. More precisely, they are distinguished by selection pressures relevant to a given explanation. Signals can be viewed as cues when the sender's functional status – whether it is coadapted with the receiver – is ignored. Similarly, signals can be viewed as influence when the receiver's functional status is ignored. The receiver is treated as part of the environment,

[12] See section 5.5 for comments on the relationship between the strategic/efficacy distinction and source/channel coding in communication theory, and section 7.4 for examples of strategic and efficacy design in nature.

and the distinction between efficacy costs and strategic costs collapses.[13]

1.5 Interaction hypercubes

The boundaries separating signals, cues, and influence are porous, with many intermediate cases. This reflects the fact that these categories are evolutionarily continuous. Cooperative signals evolve from both cues and influence. A fourth category of cooperative **mechanisms** (discussed further below) also serves as an evolutionary precursor to signals. To distinguish signals from other kinds of interaction, instead of providing necessary and sufficient conditions on signalhood, we can identify property dimensions that jointly determine a **hypercube**.

1.5.1 What are hypercubes?

Hypercubes help distinguish dimensions of variation. They allow us to represent and compare phenomena according to the different kinds of properties they possess. The term 'hypercube' in mathematics means an n-dimensional analog of a cube. Here I use it to mean a (possibly) n-dimensional space, that has the shape of an n-dimensional cube, within which different phenomena or concepts can be represented. Sometimes these spaces are only 3-dimensional (figure 1.5 A, B and D); sometimes more than three dimensions are specified but only three can be graphically represented at once (figure 1.5 C, figure 1.10). One could simply call these 'cubes', but that would miss the implication that these spaces can in principle possess more than three dimensions.

[13] I discuss **deception** at length in chapter 6.

Use of the hypercube as a means of representation has a patchy history. It seems to have been introduced several times independently. In the context of ecology, Hutchinson (1957, p. 416) defined the "fundamental niche" of an animal as an "n-dimensional hypervolume" comprising all environmental variables relevant for that creature's survival. His goal was not to depict any such volume, only to show that the concept of environmental niche was a sound one.

Churchland (1996, 24ff) uses a spatial representation to explain how sensory representations are categorised by the brain. Different kinds of taste cell (sweet, salty, sour) combine at different levels of activation to produce a three dimensional "taste space". Familiar tastes can be located within this domain (figure 1.5A).[14] Churchland depicts similar cubes for colour and face space. These representations also have a comparative use: Churchland's figure 2.4 (his page 27) depicts the relative size of human and canine olfactory spaces, starkly illuminating the extent of a dog's universe of smells. Gärdenfors (2004, 2014) tries to extend perceptual to conceptual spaces, eventually to ground a theory of semantics. Hypercubes are therefore a potential means of capturing cognitive representations.

In another domain, Mitchell (2000) argues for a continuum view of scientific laws. Comparing physical to chemical and biological 'law', she distinguishes three dimensions across which scientific generalisations can be compared: strength, stability and abstraction. Mitchell's figure 3 (her page 263) depicts several scientific laws and philosophical generalisations according to their ranking on these three scales (figure 1.5B). For example, the law of conservation of mass is extremely abstract, strong and stable. At the other end of the spectrum,

[14] Confusingly, Churchland (1996, p. 24) cites "Jean Bartoshuk" as the source of the image on his page 24. There is a psychologist named *Linda* Bartoshuk who works on the psychophysics of taste; I have not yet found a corresponding diagram in her writings.

Goodman's generalisation that all the coins in his pocket are copper is neither strong, nor stable, nor abstract. While different propositions could be ranked on a single dimension of contingency (depicted in Mitchell's figure 2, p.253), the relations between them are made much more explicit by teasing apart dissociable dimensions.

More recently, Godfrey-Smith (2009) puts the hypercube method to work by characterising Darwinian populations. Godfrey-Smith highlights five properties of populations that help determine whether they are paradigmatically, minimally, or marginally Darwinian (three of these dimensions are depicted in figure 1.5C). For example, paradigmatic cases are those with high-fidelity heredity, a smooth fitness landscape, and whose reproductive fitnesses depend on intrinsic character (Godfrey-Smith, 2009, fig 3.1 p.64). This last property contrasts with extrinsic causes of reproductive differences, such as fatal lightning strikes. Choosing five dimensions allows for mixing and matching depictions of two- and three-dimensional spaces.

Finally, Dennett (2017) extends Godfrey-Smith's account to populations of cultural items. He first places several culturally replicating populations into Darwinian space (his figure 7.4) before creating his own cube to capture different aspects of *design* (his figure 7.5) (figure 1.5D). The least 'intelligent' form of design is characterised by low comprehension, bottom-up processes and random search. Termite mounds and other animal structures belong in this corner. In the opposite corner, design resulting from top-down directed search by comprehending agents characterises many paradigmatic human cultural achievements such as Gaudí's *Sagrada Familia*, Turing's theory of computation and Bach's cantatas.

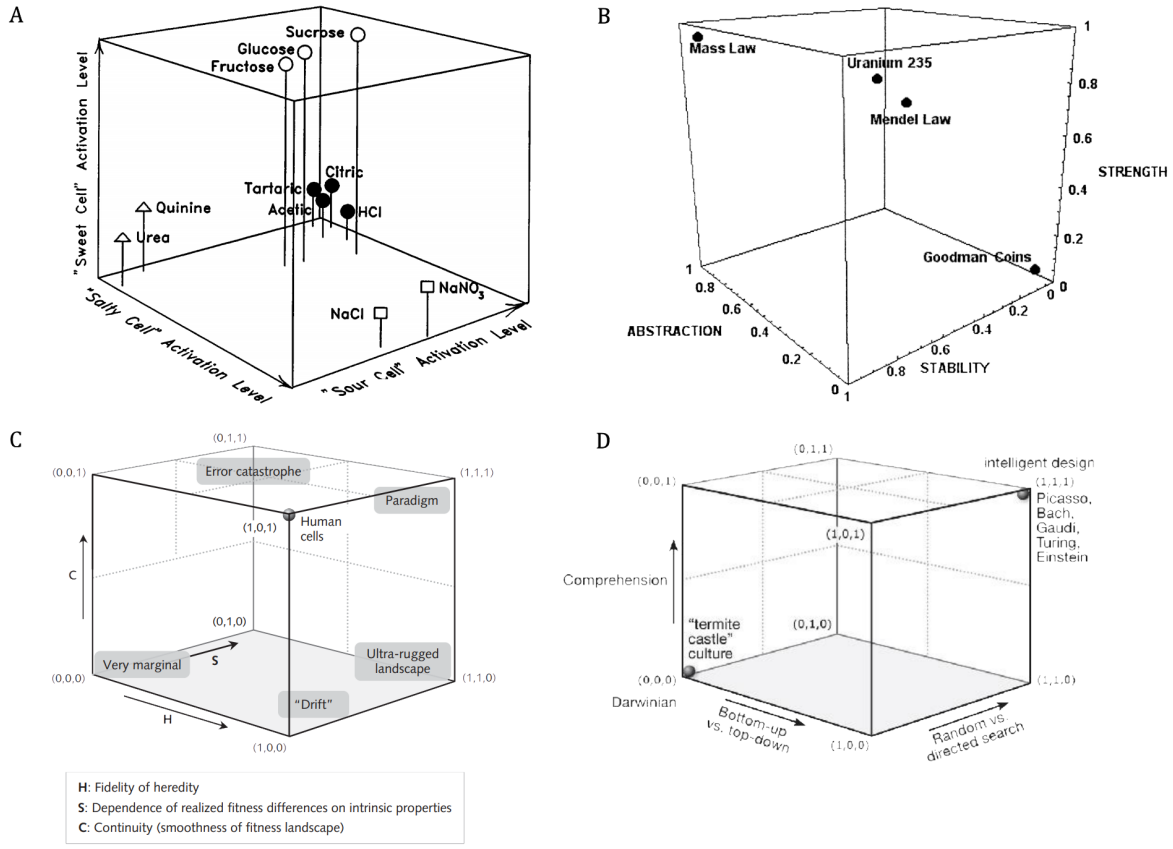

Figure 1.5: Examples of hypercubes depicted as three-dimensional cubes. **A** Churchland (1996, p. 24), "Taste space : the position of some familiar tastes." **B** Mitchell (2000, p. 263), "Multi-dimensional conceptual space for scientific laws." **C** Godfrey-Smith (2009, p. 64), spatial representation of Darwinian populations. **D** Dennett (2017, §7), "Inverted Darwinian space with Darwinian phenomena at (0,0,0) and intelligent design at (1,1,1)."

1.5.2 A hypercube for biological interactions

Signalling interactions may be characterised by several features. I have tried to select the most salient ones for my purposes. Consider the following:

- **Common interest**: the extent to which selection acted on sender and receiver jointly.
- **Resource separation**: the extent to which the material and energy required to produce the response is provided by a source other than the intermediary.
- **Arbitrariness**: the extent to which a different signal vehicle (different codes, different mapping relations) could have performed the same signalling function.
- **Transmission rate**: bits per second transmitted on average by each signal.

We can depict a maximum of three dimensions at a time. In this section I use 2D and 3D spaces. As an example, figure 1.6 depicts several interactions already mentioned in a cube defined by common interest, resource separation, and arbitrariness. I now introduce each dimension in turn.

1.5.3 Common interest

Common interest captures the extent to which sender and receiver enjoy similar benefits from similar outcomes. The intuitive way of representing common interest is as a line with full common interest at one end and conflict of interest at the other (figure 1.7A). This representation conflates influence and cue-reading, both of which entail divergent interests. There is a more subtle way of conceiving the dimension that would allow us to distinguish cues and influence. We could let the ends of the continuum represent *differential benefit* of sender and receiver. In this way, cue-reading exploitation and manipulative influence appear at either end of the spectrum, while full common interest appears in the middle (figure 1.7B).

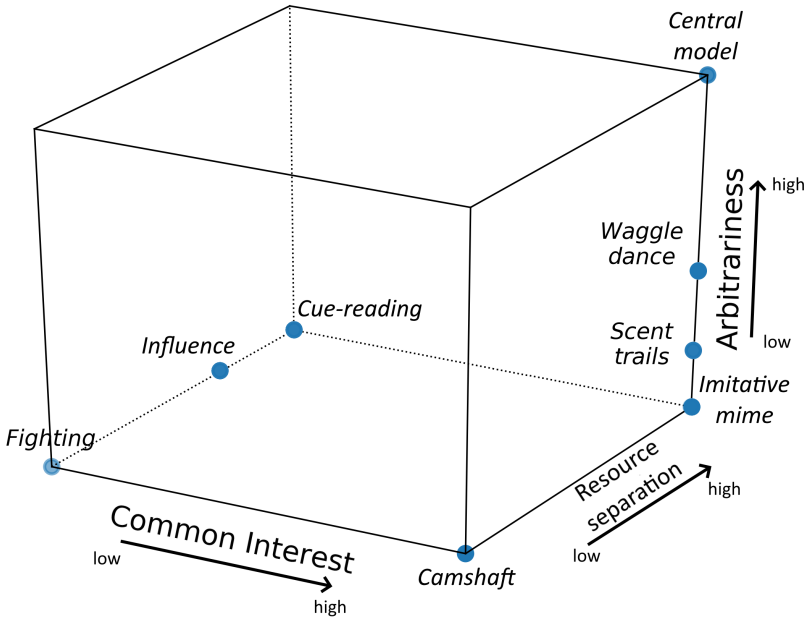

Figure 1.6: A simple example of a hypercube. The resource separation dimension captures how 'mechanistic' a phenomenon is. In this diagram and subsequently, unless otherwise indicated, dimensions should always be read as increasing from left to right and bottom to top, even where arrows are omitted.

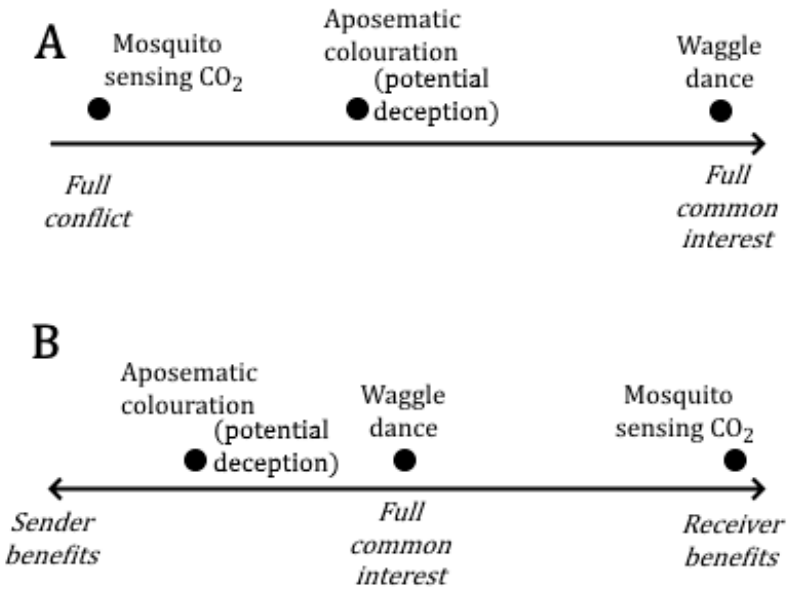

Figure 1.7: Two ways to represent common interest. **A** The canonical way. **B** An alternative.

Figure 1.7 includes three examples. One is cue-like, one is signal-like, one is influence-like. Aposematic colouration is influence-like because non-toxic senders sometimes hijack the signalling system. Treating honest and deceptive senders as the same type, senders benefit more than receivers because there is more deceit than exploitation.[15] Conversely, mosquitos sensing CO_2 gain a benefit at the expense of the mammal that emits the chemical. In figure 1.7B, eusocial behaviours like the waggle dance lie in the middle of the continuum, as senders and receivers enjoy the same benefit as a result of their close evolutionary relationship.

In the biological case, cycles may develop. When receivers become overly trusting, senders can get away with more manipulation. Deceptive senders can enter the system and

[15] See section 6.3 for considerations for and against modelling deceptive and honest signallers as the same type.

gain a benefit, pushing the system toward the left of the continuum. However, this generates selection pressure on receivers to become more discriminating. As they refuse to countenance dishonest senders, selection induces greater honesty, and the system moves to the right until it is closer to the middle.

Frank (1998, p. 247) likens these cycles to parasite/host dynamics, but notes that this perspective "has not been developed by explicit models." More recently, philosophers interested in game-theoretic representations of signalling conflict have made progress (Godfrey-Smith and Martínez, 2013; Martínez, 2019; Martínez and Godfrey-Smith, 2016; Shea et al., 2017). Building on the work of Skyrms (2010), these authors explore communication in explicitly conflictual situations.

When combining common interest with other dimensions, I exclusively use the canonical format. There are two main reasons for this. First, a reader seeing this dimension for the first time would most likely expect the canonical format. It is worth keeping to reader expectations when possible, to avoid unnecessary confusion. Second, common interest between two parties may be known before those parties engage in a signalling (or signal-like) interaction. In such a case there could be determinate conflict of interest without it yet being determined which party would benefit from interaction between the two. It may therefore be possible to place the pair on the scale depicted in figure 1.7A without it having a determinate spot in figure 1.7B: as it were, there are two positions on figure 1.7B compatible with each position on figure 1.7A (with the exception of full common interest).

1.5.4 Resource separation

> There is in electrical engineering a split which is known in Germany as the split between the technique of strong currents and the technique of weak currents, and which we know as the distinction between power and communication engineering.
>
> Wiener (1948, p. 39)

François Jacob (1973, p. 251), in a treatise on molecular systems, spoke of two forms of power: "the power to do and the power to direct what is done." Paradigmatic signals trade away the first in order to capitalise on the second. They guide receivers in performing actions. The material and energy for those actions is provided from elsewhere. In other words, paradigmatic signals are those for which energy resources used to send the signal, and material resources from which it is made, are distinct from those energetic and material resources used in the receiver response. We can call this aspect *resource separation*.

This dimension, while clearly significant, is curiously underappreciated in the literature.[16] We must at least try to capture the intuition, even if not formally. Here is my suggestion.[17] Consider a putative signal. Take the energy and material that the receiver uses to perform the response, and calculate the percentage of those resources provided by the signal itself. A high percentage means that most of the energy and material used to perform the response was provided by the putative signal. This makes for low resource separation. On the other hand, if little or none of the resources consumed in the response derive from the

[16] There are at least two significant exceptions. First, in the context of animal communication, Wiley (1994, p. 162) defines a signal as "any pattern of energy or matter produced by one individual (the signaler) and altering some property of another (the receiver) *without providing the power to produce the entire response*" (emphasis original). Second, since the original version of this thesis was submitted, Artiga (2020) has defended a claim very similar to that suggested here. See also footnote 2 page 11.

[17] I am very grateful to the anonymous examiners of this thesis for refining and clarifying this suggestion.

signal itself, there is high resource separation.

Paradigmatic signals have high resource separation. All of the energy and material required to produce the response is provided by the receiver (or some other source that is distinct from the signal). Signals in the central model can be classed as paradigmatic. While it is typically left unspecified where the energy to perform the response (decoding) comes from, it is assumed that the energy required to send the signal is just used to overcome noise. It is implicit that the receiver has its own energy source, and therefore that none of the energy required to produce the response is provided by the signal itself.

At the other extreme, some intermediaries between putative senders and receivers are such that all of the energy required to produce the response is provided by the 'signal'. Such intermediaries are not typically considered signals at all. For example, a camshaft in a combustion engine shares several properties with paradigmatic signals. It mediates between a sender and receiver. The shape of the cam and the rotational speed of the shaft determine the rate at which a valve opens and closes. Since the valve's proper function is to open and close at a certain rate, it counts as a receiver guided by the cam. Camshafts fit our preliminary characterisation of signalling behaviour. We *could* think of the cam as 'directing' the valve to open and close, but we do not *want* to think this. The valve does not 'interpret' the position of the cam; rather it is physically pushed by it. All of the energy needed to open the valve is provided by the cam. There is no resource separation in this system: the receiver's act is a certain kind of regular movement, and all of the energy required to perform this movement is provided by the intermediary.

Signals can evolve along the dimension of resource separation. Contemporary relatives of the Venus flytrap use sensitive tentacles to wrap around prey, limiting their escape (see the

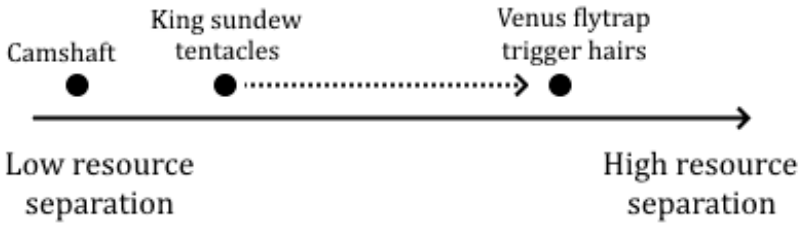

Figure 1.8: Paradigmatic signals are those for which the resources required to produce the response are separate from the signal itself. The camshaft causes an instant effect, but all of the energy of the effect is provided by the camshaft itself. Tentacles of the King sundew (*Drosera regia*) slowly wrap around prey, causing leaves to enfold the victim. We might suppose that leaves move partly under their own power and are partly 'pulled' by the tentacles. Finally, the trigger hairs of the Venus flytrap (*Dionaea muscipula*) are a specialised version of those tentacles, whose only job is to trigger the snap trap shut. The energy required to shut the trap is stored as elastic potential of the trap, rather than being provided by the action potential. Evolution of the snap trap from a *Drosera*-like ancestor corresponds to movement from left to right along this dimension.

case study on p.42). Similar tentacles were likely possessed by a common ancestor with the flytrap, and evolved into its trigger hairs. Figure 1.8 depicts this relationship.

How the Venus flytrap snaps

The Venus flytrap (*Dionaea muscipula*) famously catches insect prey in a rapidly shutting chamber called a snap trap. Nectar and a brightly coloured surface attract unwary insects to a platform, from which sprouts a bank of trigger hairs. When prey stumble over the hairs, action potential occurs. The electrical signal causes a change in the curvature of the open leaf, leading to a rapid release of elastic potential energy (Forterre et al., 2005). This energy release is constrained by the geometry of the leaf, and the trap snaps shut.

Snap traps evolved only once. Two related genera use them, with one species in each. The Venus flytrap represents *Dionaea* while the waterwheel plant (*Aldrovanda vesiculosa*) is the only member of *Aldrovanda*. The closest related genus is *Drosera*, which contains many sticky-trap plants. We tend to look there for pre-adaptations and clues to evolutionary history. For example, tentacles of the king sundew (*Drosera regia*) slowly enfold the plant's victim, causing leaves to wrap around it. Eventually leaves fully enfold the prey, allowing tentacles to digest extracted nutrients. The trigger hairs of the Venus flytrap are a specialised version of those tentacles, whose only job is to trigger the snap trap shut. Gibson and Waller (2009, p. 581) describe *Drosera* tentacles as "an essential 'pre-adaptation' for (and homolog of) the trigger hairs" of the Venus flytrap. These tentacles became trigger hairs in the Venus lineage.

Snap traps seemingly allow larger prey to be captured more reliably. Tentacles become teeth and trigger hairs by losing their stalks and mucilage, the gluey substance used to trap prey (Snyder, 1985; Williams, 1976). In other words, tentacles lost mechanical features as they gained a signalling function. In plants, action potentials usually have an osmotic rather than communicative purpose. That the same signalling mechanism arose in Venus flytraps (and other vascular plants such as *Mimosa pudica*) and in metazoan nervous systems is an example of convergent evolution.

1.5.5 Arbitrariness

The notion of arbitrariness is notoriously multifaceted. We can draw on Planer and Kalkman (2019) who distinguish two definitions of an arbitrary signal. Roughly these are signals that

do not resemble the signified and signals that *could have been different*.

First, a signal is arbitrary$_R$ if it does not resemble its signified. This means it is not *iconic*. At first glance this looks like a useful dimension for our task. However, Planer and Kalkman employ a strong sense of "iconic", one which requires appreciable cognitive sophistication and rules out (for example) the waggle dance as iconic. *Every* signal discussed in this thesis is arbitrary$_R$, which means the notion cannot do any work for us (though it does plenty of good work for Planer and Kalkman). In the following, I leave arbitrariness$_R$ aside.

In contrast, arbitrariness$_A$ is just what we need. A signal is arbitrary$_A$ if structurally dissimilar signals might have played the same communicative role. Modality is doing all the work here: how we spell out "might" determines which signals count as arbitrary in this sense. For example, bacteria use acyl homoserine lactone (AHL) to coordinate quorum sensing (Hughes and Sperandio, 2008). Other signalling molecules are known to play similar roles, such as autoinducers AI-1, AI-2, and CAI-1 (Mehta et al., 2009, p. 2). If those alternatives could play the same chemical roles as AHL, the actual AHL signal would be somewhat arbitrary. This is consistent with AHL having several signalling roles, for some of which it is more arbitrary than others.

From now on, unless explicitly stated, the term *arbitrary* should be read as meaning arbitrary$_A$.

The first piece of good news is that this definition accords with intuition that signals in the central model are extremely arbitrary, and most biological signals are not. Signals in the central model are encoded sequences of 1s and 0s. There are a wide variety of different codes available for different engineering purposes, and even within the same coding scheme there are typically many different ways to assign codewords to source symbols. Any given choice

of codeword assignment is clearly arbitrary. In contrast, biological signals are typically tightly bound to the sensory and behavioural constraints of the agents involved. Planer and Kalkman (2019, pp. 15–6) discuss the kinds of internal and external constraints that impact the size of the set of viable alternative signals in a given context.

More good news: the account allows us to distinguish between the arbitrariness of different *aspects* of signals. Words of human natural language are the canonical example of arbitrary signals, but they are decidedly *not* arbitrary with respect to their length. Indeed, a rather robust generalisation called Zipf's law ties a word's length to its frequency of use (Kanwal et al., 2017). Plausibly, frequency of use has some relation to semantic content. Therefore, at least with respect to this aspect, word form is not so arbitrary after all. This doesn't threaten the view of human language as paradigmatically arbitrary, because words possess many other features. For example, different phonetic combinations could in principle serve the same purpose for a given word.

The canonical example of an arbitrary biological signal is the genetic code. Without going into detail, the translation between nucleotide triplets and amino acids is thought to be contingent: different assignments would work just as well in converting nucleotide sequences into proteins. In this regard, Stegmann (2004) is the usual work cited; although it is unclear how similar his definition is to the notion of arbitrariness we have adopted, it should be fairly obvious that the code is arbitrary in the sense we want. However, as with language, there may be unnoticed efficiency disparities between rival codes. Such effects would threaten the arbitrary status of the code (Hofstadter, 1985, §27).

See again figure 1.6 for examples of familiar signals that differ along this and other dimensions.

1.5.6 Transmission rate

As presented in the central model, transmission rate measures the number of bits per signal.[18] Bits are measures of probability change. Transmission rate is therefore a measure of probability change per signal, or more transparently, a measure of how much signals change probabilities.[19]

For example, honeybees improve foraging success by restricting their search space to locations indicated by the waggle dance. Each dance changes the probability distribution of nearby food in a principled way. From this perspective, transmission rate measures the accuracy of a signal. The greater the change in probabilities, the more reliably receiver bees find the indicated food source. Section 2.2 explores in more detail informational measurements of the waggle dance derived by Haldane and Spurway (1954) from Karl von Frisch's statistical data.

Most biological signalling channels have low bitrate, especially compared to artificial channels. This is because the range of choices biological devices face is much lower than that of human-designed systems. A computer monitor must choose between billions of possible configurations each second, requiring Megabit transmission rates from the processor. In contrast, biological devices (perhaps excluding those associated with cognition, such as perceptual states) typically choose among a small class of options.

[18] When the number of signals that can be sent per second is known, this entails a rate given in bits per second.

[19] Recent work on the notion of *causal specificity* (Griffiths et al., 2015) suggests measuring the extent to which a cause is specific to its effect as the mutual information between them (strictly, between the effect and *interventions on* the cause). This characterisation is analogous to the transmission rate concept, which is measured as the mutual information between transmitted and received signals (appendix A). Part of the benefit of this concept of causal specificity is that it allows us to precisify the idea that genes are more specific causes of their products than non-genetic mechanisms. This perspective aligns with my claim (below) that genetic transcription and translation possess relatively high transmission rates.

Choices need not be between discrete options. Consider a biological device facing the problem of how much of a particular chemical to release. Distinctions between different volumes are not discrete, and should be modelled using continuous parameters. The waggle dance does not distinguish discrete partitions of the surrounding environment, and it too should be modelled continuously in order to derive the appropriate transmission rate (see chapter 2).

Adding a dimension for transmission rate to our hypercube provides the most obvious distinction between artificial and biological signals (figure 1.9). While artificial devices are strictly cooperative and typically high-rate, biological signals are beset by divergent interests and their transmission rate is often very low.

The genetic code deserves special mention. If we restrict attention to the operation of transcription and translation machinery inside the cell, the genome is an instruction for building a sequence of amino acids. We can model the situation as follows. There are 64 possible messages (4 nucleobases, 3 bases per codon, $4^3 = 64$) and 21 possible outcomes (20 amino acids plus one 'stop' action). Maximum transmission rate would occur when all outcomes are equiprobable, so each codon would transmit $\log 21 = 4.39$ bits.[20] According to Fuchs et al. (2014) translation is slower than transcription, so we need only look to the former to find the maximum available rate. Prabhakar et al. (2017, p. 1354) cite Young and Bremer (1976) as observing a translation rate of 5-20 amino acids per second in *Escherichia coli*. Each amino acid is indicated by a single codon, so the maximum rate would be 4.39 bits/codon or 87.8 bits/second. Compare this to the honeybee waggle dance. Haldane and Spurway

[20]The *capacity* is larger than this because (ignoring noise) each codon could carry $\log 64 = 6$ bits by signifying a unique outcome.

(1954, p. 255) estimate each dance to carry at most 5 bits about the location of food (see also section 2.2). The duration of each dance is highly variable and dependent on factors such as food quality. Nonetheless, we can take the results of Seeley et al. (2000, Table 1, p.815) to indicate a very rough average of 18 seconds per dance.[21] Such a dance conveys around 5/18 = 0.28 bits/second. Although the axis of figure 1.9 is not to any kind of scale (and the central model should be thought of as *indefinitely* higher than the genetic code), the disparity between the waggle dance and the genetic code reflects these calculations.[22]

1.5.7 Placing paradigms on the hypercube

Each scientific discipline draws on proprietary explanatory resources. Historical contingencies have led to different subdisciplines covering different regions of conceptual space. These regions are often disconnected from each other, but they are in principle continuous. After all, one of the lessons of evolutionary biology is that discrete typological thinking is often inappropriate. Already well-learned in the case of species and other groupings, the lesson holds true for behavioural interactions too.

We can place entire disciplines on the hypercube, based on what they assume about signalling (figures 1.10, 1.11, 1.12, and 1.13):

- **Cellular signalling**: fully cooperative, high resource separation, not very arbitrary, low transmission rate.

[21] This time was noted for a food source consisting of 2.50 mol/litre sucrose solution, and includes only time spent dancing.

[22] An anonymous examiner suggested using bits/channel use rather than bits/second to measure this dimension. One of the problems with a bits/channel use measure is determining what counts as a single use of the channel. For the genetic code, is each codon triplet a single channel use? Or each strand of DNA? If the latter, channel uses can vary greatly in length. Furthermore, a telegraph wire might only have 1 bit/channel use (because each binary symbol counts as a single use of the channel), while the waggle dance has 5 bits per dance. This fails to capture the vastly greater information rate that can really be transmitted by telegraph.

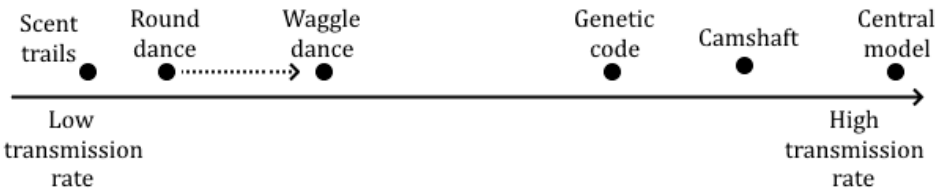

Figure 1.9: Transmission rate (bits/second) in the **central model** is determined by the size of the lexicon and properties of the channel, and is in principle unlimited. In comparison, the transmission rates of biological signals are extremely small (with the notable exception of the **genetic code**; see main text for details). **Scent trails** appear to have similar bits/signal measures to the waggle dance (Wilson, 1962) but take longer to traverse on average than the time it takes to perform a waggle dance, so scent trails transmit fewer bits/second. The **waggle dance** itself probably evolved from a less precise (and therefore lower-rate, see section 2.2) communicative behaviour called the **round dance**. Notice that mechanisms like the **camshaft** would have transmission rates if treated as putative signals; they 'select among' causal effects. Transmission of a single bit would be interpreted as the opening or closing of the receiver valve, as those are the two functional states controlled by the camshaft. The rate of a camshaft would be determined by converting revolutions/minute into bits/second, and would be relatively high in comparison to biological signals.

- **Bacterial signalling**: largely cooperative, high resource separation, not very arbitrary[23], low transmission rate.
- **The genetic code**: largely cooperative, high resource separation, extremely arbitrary, high transmission rate.
- **Animal signals**: conflictual, high resource separation, large variation in arbitrariness, low transmission rate.
- **Central model**: fully cooperative, full resource separation, very arbitrary, high transmission rate (in principle unbounded).

Two distinct evolutionary origins of signals can be understood as two paths within the cube. First, a path leads from cues and influence on the bottom-left edge of figure 1.10 (or bottom-left corner of figure 1.11) to the various forms of biological signalling on the right. This is the pathway cited in discussions of ritualisation, and inter-organismal signalling appears to be very often a result of it. It can be further divided into two, depending whether the original interaction was cue-like or influence-like. Second, a path leads from mechanical features at the front-bottom-right corner to certain kinds of intraorganismal signalling (not shown). In particular, the evolution of Venus fly trap trigger hairs took this pathway. It is not clear to me how common this pathway is; especially since cellular signalling (which would be a good candidate for it) is more likely a consequence of the cue-pathway during the onset of true multicellularity. Nonetheless, it is worth pointing out this mechanical pathway as a possible explanation for the emergence of signals. Both pathways are depicted as large black arrows in figure 1.13.

[23] Both cellular and bacterial signalling might be more arbitrary than I credit here, especially given plasticity in the definition of arbitrariness$_A$.

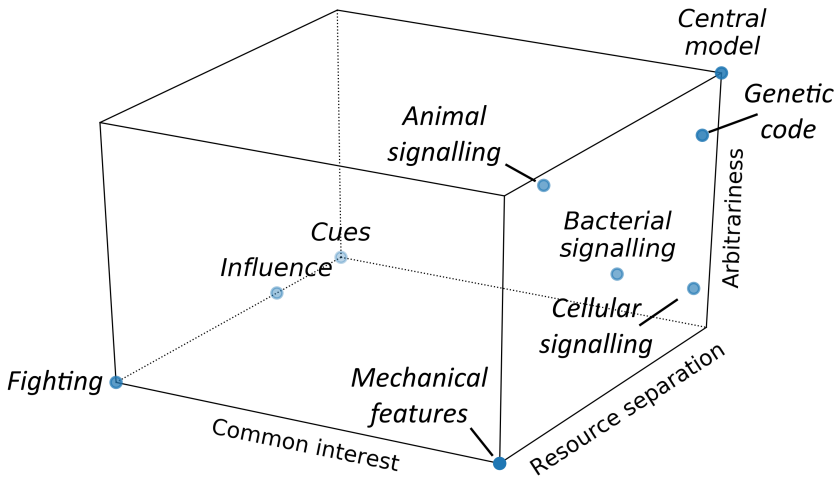

Figure 1.10: Paradigms of biological signalling, the central model of communication theory, and other non-signalling interactions. Figures 1.11, 1.12, and 1.13 respectively depict views through the front, right, and top faces of this cube.

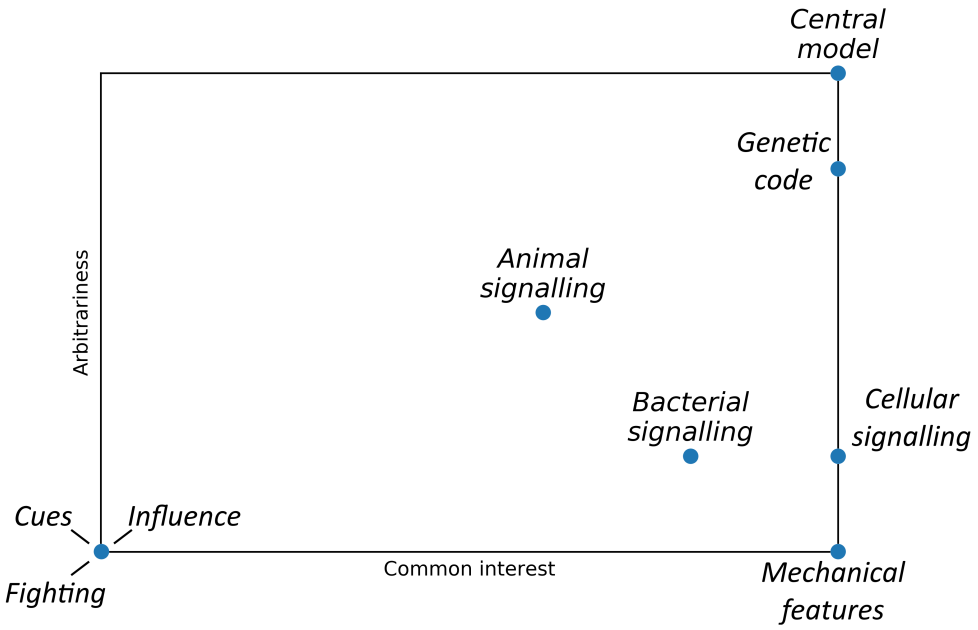

Figure 1.11: View through the front face of the cube depicted in figure 1.10.

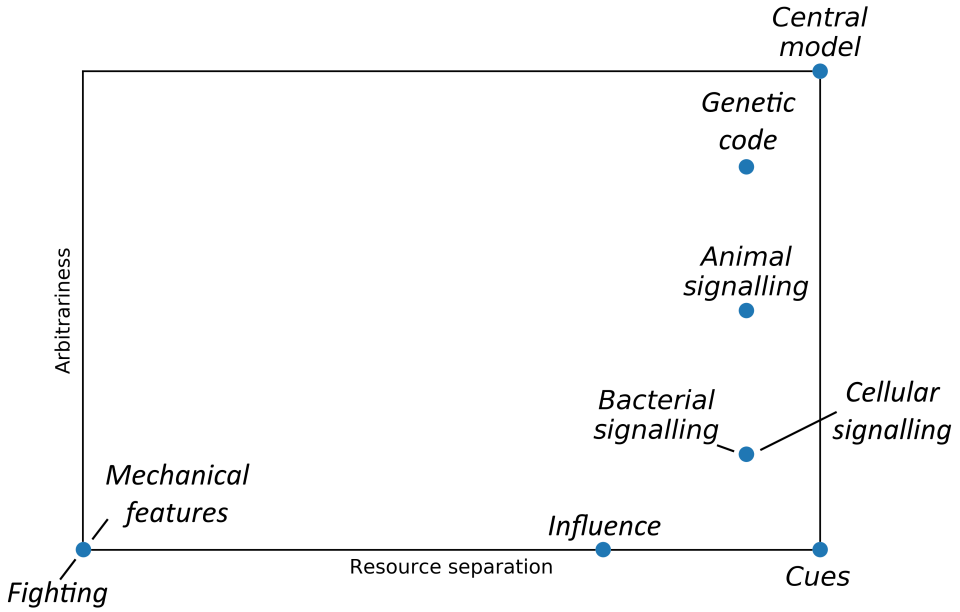

Figure 1.12: View through the right-side face of the cube depicted in figure 1.10.

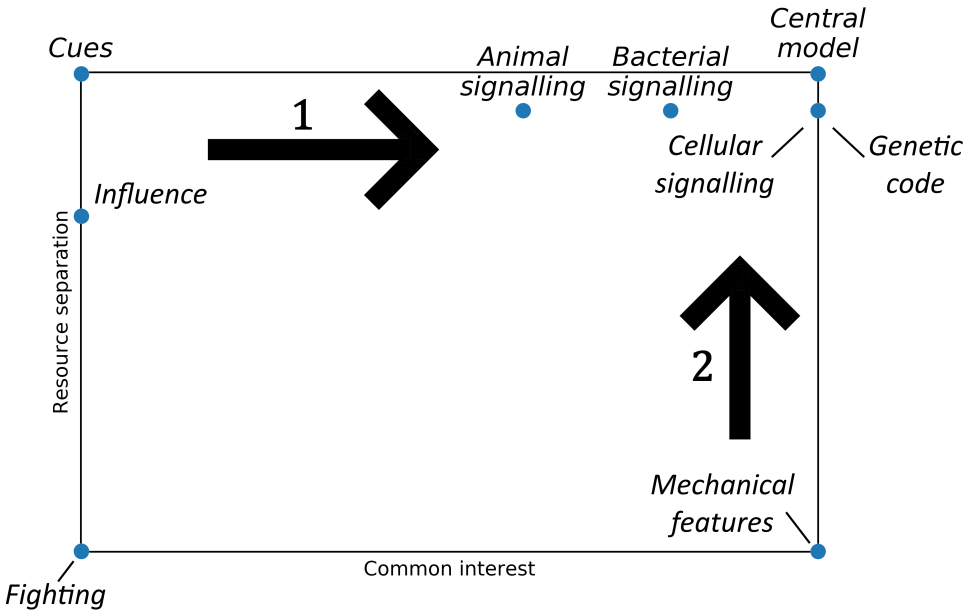

Figure 1.13: View through the top face of the cube depicted in figure 1.10 (i.e. a bird's-eye view of the cube). Two pathways of signal evolution are depicted as movement toward the region of paradigmatic signalling in the top-right quadrant: **1** evolution from cues and influence (e.g. the waggle dance); **2** evolution from mechanical features (e.g. Venus fly trap trigger hairs).

The question whether communication theory generalises to biological signals depends on whether its formal tools are restricted to the top-right corner of this cube. Note first the proximity of the genetic code to communication theory. The genetic code displays several marks of being well-engineered for its informational job (Bergstrom and Rosvall, 2011). In part because it inhabits the same area of the hypercube, Bergstrom and Rosvall (2011) argue for the applicability of definitions and concepts from communication theory in genetics.

In the remainder of the thesis, I argue those tools are generalisable across the entire hypercube. Communication theory can and should be expanded to include all phenomena under discussion. Teleosemantics provides an interpretation of communication-theoretic mathematics throughout the space.

It might be thought that game theory covers the entire space already. Indeed, what I have in mind is less an extension of communication theory than a plea for unification. Game theory, communication theory and decision theory all draw on the same formal repertoire. For some reason, many scholars accept game theory and decision theory as modelling strategies in biology, while disbarring communication theory from the same job. By the end of the thesis, it will be clear that such a position is inconsistent.

1.5.8 Marginal cases and non-cases

Teleosemantics is thought by some to be too liberal because it includes mechanistic intermediaries within its scope. However, the central insight behind the theory is made more plausible by focusing on paradigmatic signals. Hypercubes are one particularly vivid way of distinguishing paradigmatic signals from marginal cases.

The hypercube analysis shows that the sender-receiver paradigm allows marginal cases.

This claim had previously been disputed by Neander (2017). Neander argues that sender-receiver accounts impose all-or-nothing requirements on putative signals (Neander, 2017, p. 277). She argues from this premise to the conclusion that sender-receiver accounts attribute content to vehicles that intuitively are contentless. The example she draws on is a class of enzymes with the generic name protease, used in the digestive system. Protease is transported from the pancreas to the gut to aid in digestion. Neander takes this system to have sender-intermediary-receiver structure. According to sender-receiver teleosemantics, protease would therefore counts as an intermediary between cooperating entities. Furthermore, the gut can digest protein when protease arrives, and increases in protein are accompanied by increases in the release of protease. Given these facts, the sender-receiver account entails that protease is a signal of the presence of protein in the gut. Neander finds this result implausible and cites Sterelny (1995, p. 256), Ramsey (2007), Burge (2010, pp. 294-307) and Schulte (2015, p. 10) in her defence. Rejecting protease as a signal (she uses the term 'representation'), Neander treats it as a counterexample to sender-receiver teleosemantics.

However, there are two problems with Neander's argument. First, she claims that sender-receiver teleosemantics places "all or nothing requirements" on representational vehicles (Neander, 2017). In this section I have motivated a view to the contrary. Sender-receiver teleosemantics, like any theory that purports to outline a kind with which scientists can theorise, allows for marginal cases. When it comes to signals, I have tried to outline where I think the margins lie: primarily at the limits of common interest and resource separation, secondarily at the limits of arbitrariness and transmission rate. If protease really is an intermediary in a sender-receiver partnership, it might not qualify as a paradigmatic signal due to failing to lie within the paradigmatic region of the resource-separation dimension. From the

facts presented so far, it does not follow that sender-receiver teleosemantics must attribute signalhood to protease.

Second, when we investigate the mechanical details of protease, we find a more serious challenge to Neander's claim. Neander considers the gut (i.e. the small intestine) to be the receiver of the protease signal. However, the way a protease such as trypsin works suggests that the system does not have the right kind of sender-receiver structure. Trypsin catalyses the process by which proteins are broken down. It does this *within* the small intestine after being transported there from the pancreas. It does *not* have a causal effect on the small intestine that then prompts the intestine itself (or some part of it) to catalyse the reaction. The intestine is simply the place where the reaction occurs, because that is where the relevant proteins are found during digestion.

It is worth noting that my response would not be plausible if the reasons why sender-receiver teleosemantics rejects protease as a paradigmatic signal were *different* than the reasons why biologists reject it as such. For then it would be a lucky accident that teleosemantics gets the right answer, rather than as a result of correctly describing the concept (or kind) 'signal'. In this case, they are the same reasons: a protease such as trypsin is not treated as a signal because the proteins on which it acts are taken to be resources for an organism rather than constitutive parts of it. If they were the latter, proteins would perhaps be ascribed the function of breaking down into smaller parts; consequently, catalysis would perhaps be described as the triggering of this function; consequently, trypsin would perhaps be considered a signal molecule. It is because the molecules whose breakdown is catalysed by enzymes are *not* considered functional parts of an organism that enzymes are not considered intermediaries, and are therefore not even putative signals.

There is a problem with the response I have just offered to Neander's objection. The protease system is even more complex than I have so far allowed. Trypsin is stored in the pancreas as trypsinogen, an inactive form that protects the pancreatic tissue itself from the effects of trypsin. Trypsinogen is transported to the duodenum (part of the small intestine) in this inactive form. There, an enzyme called enteropeptidase is secreted, which converts tripsinogen into trypsin. So there is a sense in which a causal interaction between the intermediary (trypsinogen) and the receiver (enteropeptidase, broadly construed as part of the gut) effects a response (trypsin production). We might squeeze this into a sender-receiver template after all.

However, the receiver's act in this case is the production of an enzyme (trypsin) from materials already present in the putative signal (trypsinogen). Here is where the paradigm/-marginal structure I have been arguing for kicks in. Most or all of the material required to produce the response is already present in the putative signal. It is therefore *not* a paradigmatic signal, because it scores very low on the dimension of resource separation.

Again, it is not a lucky accident that sender-receiver teleosemantics gets the right result once we look at the case in more detail. The hypercube analysis captures those features of causal intermediaries that biologists consider significant when ascribing signalhood. Suppose the molecule transported from the pancreas to the duodenum were not trypsinogen (i.e. not a molecule that contained the constituents of trypsin), but a molecule that stimulated the production of trypsin by the duodenum. Then, I submit, biologists would consider that transported molecule a signal, and the pancreas and the gut would have sender-receiver structure as required. Teleosemantics would assign content equivalent to 'there is protein in the gut; produce trypsin!' to the transported molecule, in agreement with the biologists' own account.

Teleosemantics, together with the hypercube analysis, captures significant aspects of biologists' practice of ascribing signals. That is why it gets the right answer in these actual and hypothetical cases.

1.6 Why teleosemantics?

There are plenty of philosophical accounts of meaning. What makes sender-receiver teleosemantics uniquely well-suited to interpreting biological communication?

Ideally I would argue that teleosemantics is the best account of meaning available, but that is not the focus of the present work. Strictly, this thesis is about applying communication theory in biology. Teleosemantics is an interpretive framework that resolves philosophical problems relating to that application. There may well be other frameworks that resolve these problems as powerfully. However, I find it very plausible that a comparable result would be forced to include function as part of the analysis. The best account will be a teleosemantic theory of one sort or another. With that in mind, I will restrain myself to comments in favour of Millikan's theory over other versions of teleosemantics.

Millikan's theory captures correctness conditions within the central model itself (chapter 3). This allows us to starkly demonstrate the relevance of communication theory for explanations of signalling behaviour. Two alternative versions of teleosemantics, due to Neander (2017) and Shea (2018), put distance between semantic content, characterised by correctness conditions, and the concepts and tools of communication theory.

Neander offers her own definition of information, with the hope of using it to construct a workable notion of semantic content for cognitive science. Neander (2017, p. 142) defines

information as follows: "one event carries information about another if the second causes the first" (emphasis removed). On the face of it, this definition looks simultaneously too liberal *and* too restrictive. It is too liberal because A can cause B without B carrying information about A: there may be other causes of B that make it impossible to infer anything about A. It is too restrictive because it excludes exploitable correlations: non-causal statistical relationships that persist for good reason. It also violates symmetry: typically, definitions of information allow that if A carries information about B then B carries information about A.[24]

Furthermore, Neander gives short shrift to communication theory. In describing the central model, she says:

> Shannon offers no analysis of the relation in virtue of which a sign carries information about a state of affairs (his interest was in other issues). And, while information theory in the tradition of Shannon's work remains an important resource for the mind and brain sciences, it is an open question to what extent the notion of information used in these sciences – in talk of the flow of information in the brain and the like – is constrained by this tradition.
>
> Neander (2017, p. 7)

To the contrary: biology, and by extension cognitive science, should be directly constrained by that tradition. Coding strategies are precisely concerned with how signals correspond to signifieds. Indeed, coding strategies are arguably more useful to communications engineers than the theorems that made Shannon famous. The source coding theorem, for example, guarantees that an optimal code exists in principle for arbitrarily long messages (see appendix A). But knowing *which* code to use, given that messages are not arbitrarily long in practice, is

[24] Neander (2017, n. 19, p. 266) considers this problem and suggests introducing two kinds of information (one for each temporal direction); however, she later raises further problems with this suggestion that remain unresolved (Neander, 2017, n. 28, p. 268).

much more useful. And, as I argue in chapter 3, knowing which code to use *just is* knowing how to imbue a signal with the information it carries about the source.

Next, consider Shea's varitel semantics (Shea, 2018). Shea draws on the familiar mathematical definition of information. But, like Neander, he bemoans the limitations of communication theory:

> The 'information' of information-processing psychology is a matter of correctness conditions or satisfaction conditions, something richer than the correlational information of information theory. [...] Correlational information may well be an ingredient in a theory of content [...] but even the sophisticated tools of mathematical information theory are not enough, without other ingredients, to capture the core explanatory difference between correct representation and misrepresentation.
>
> Shea (2018, pp. 12–13)

Along with Martínez (2019, 2020) I argue that this claim is slightly too quick. There *is* an explanatorily relevant notion of failure in engineering systems, described by the mathematics of communication theory, that corresponds with the teleosemantic definition of misrepresentation (see chapter 3). The formal work that best illustrates this point is a relatively obscure corner of the mathematics called *rate-distortion theory*. Sender-receiver teleosemantics gives clues to the proper interpretation of these unfamiliar results (see chapter 6).

In sum, by providing an abstract characterisation of signalling that captures the central model of communication theory, Millikan's teleosemantics paves the way for applications of communication-theoretic tools and concepts in biology. Other versions of teleosemantics explicitly disavow that the connection is as tight as I claim. Chapter 3 offers further arguments in support of my position.

1.7 Conclusion

Interactions between biological agents can be conceived as belonging to a multidimensional space. Paradigmatic signals are interactions that take place within a particular region of the relevant hypercube. For signals of this kind, informational measurements can be made by borrowing models from communication theory. Sender-receiver teleosemantics is particularly well suited to interpreting these models and measurements.

Chapter 2

Measuring information in a biological signal

Parts of this chapter are adapted from Mann (2018).

2.1 Introduction

This chapter establishes grounds on which attributions of information in animal signals are warranted. Informational language is most appropriate when signalling behavior is the result of coadaptation. Several lines of argument are presented in favor of the use of information theory for studying certain biological communication systems. Throughout, a positive proposal is advanced for the interpretation of information in biological signals. Information is a measure of the accuracy with which a shared goal is achieved. In our case study, which focuses on the waggle dance of the honeybee *Apis mellifera*, the goal is finding food. Insofar as receiver bees have a greater probability of finding food after having followed a dance, information is transmitted by dancing bees. It is this notion of transmission that constitutes grounds for application of the information concept, and related mathematical tools, in studies of biological signalling.

The communication theory approach is a special case that emerges as sender and receiver interests become aligned. These considerations give grounds for optimism about the use of informational concepts and measurements in behavioral ecology and biology more broadly. Paradigmatic cases of information transmission are those in which the form of the signal is

designed by coadaptation of sender and receiver. In these cases, signal structure can appropriately described in terms of a 'code'.

The chapter is structured as follows. In section 2.2 I describe a classic study of the honeybee waggle dance in informational terms. I explain and interpret the model and equations used in that study, as well as making explicit the assumptions required for those interpretations to hold. In section 2.3 I respond to two preliminary objections to the model. I show that both are based on mathematical and/or interpretive mistakes. In section 2.4 I discuss a more general family of objections to the use of informational concepts in behavioural ecology. These objections cannot be sustained for all instances of animal communication, because the honeybee case shows that well-defined notions of information and encoding are applicable in at least one case. However, the informational perspective crucially rests on an assumption of cooperation between organisms. The question what becomes of information when interests diverge is picked up in chapter 7.

2.2 Measuring information in a biological signal

2.2.1 The honeybee waggle dance

"The" honeybee waggle dance is rather a family of communicative behaviours performed by all seven species of honeybee (Beekman et al., 2015, p. 1). Six species are native to South East Asia, but the seventh – the Western honeybee (*Apis mellifera*) – is widespread in Europe, Africa, and Asia and has been domesticated since antiquity, making it uniquely amenable to scientific study. Western honeybees perform the waggle dance when foraging and searching for new nest sites. The works discussed here focus on *A. mellifera* foraging at

artificial feeders and performing the waggle dance on a vertical surface inside the hive (von Frisch, 1950, pp. 76–7). The term "honeybee waggle dance" refers to this context. See the case study for details (p. 64).

The waggle dance

When individual bees discover a valuable food source, it is often beneficial to recruit nestmates as soon as possible. Competition from other foragers, and the brief duration of flowering, set important time constraints. A bee that finds a good flower patch will often try to direct available workers to that location. Indicating direction and distance on the vertical inner surface of a pitch-black hive is no mean feat. Bees nonetheless succeed by performing repeated patterns of figure-eight movement whose detectable features – orientation and duration – correspond to the direction and distance of the food source (see figure 2.1). This is the waggle dance, so called because the bee's body vibrates during the straight portion of the figure-eight run with a frequency corresponding to the quality of food discovered. Interested recruits physically follow the dancer to familiarize themselves with the direction and distance indicated, before flying off to locate the relevant food patch.

It is unclear how useful the dance is across different contexts, and the purpose for which it originally evolved. Several lines of evidence count in favour of nest site selection as the significant factor (Beekman et al., 2008; I'Anson Price and Grüter, 2015). It is relatively difficult for cavity-nesting species to find viable housing. Flowers and trees actively advertise their wares; small gaps in rocks or bark do not. In addition, nest site selection takes place under extreme time pressure, typically when

a new swarm has left its original colony and is temporarily exposed to the elements and predators. Foraging, in contrast, rewards individual search handsomely. Only in exceptional circumstances does it demand social coordination.

These arguments raise general questions about the benefits of worker recruitment over individual search (Grüter et al., 2008; Grüter and Ratnieks, 2011; I'Anson Price and Grüter, 2015) (Schürch et al., 2016, pp. 6–15). Many factors affect the relative utility of the two strategies, and it is not yet clear which cues bees use in deciding to switch (I'Anson Price et al., 2019). Since recent work on insect cognition reveals that nominally sophisticated behaviour can be reliably achieved with modest brains (Cabirol et al., 2018; Plath et al., 2017), it is difficult to confidently infer cognitive explanations for individual and social behaviour. All these considerations are salient for the general question of colony intelligence as a form of emergent behaviour. For example, a counterpart to the waggle dance is the 'stop signal', when bees physically butt scouts and cause them to cease dancing (Schürch et al., 2016, pp. 54–9). The interplay of positive and negative feedback through combined dances and stop signals is thought to ground sophisticated decision-making capacities at the level of the entire colony.

Informally, the dance is a signal about the location of food. behavioural ecologists advert to the correspondence between dance and food when explaining foraging behaviour of hive recruits (von Frisch, 1950, p. 78) (Riley et al., 2005, p. 205) (Biesmeijer and Seeley, 2005, p. 133) (Beekman et al., 2015, p. 1). Indeed, much work on honeybees after von Frisch was geared towards demonstrating that these explanations were correct (Gould, 1975). The rules of the dance language play a significant role in explanations of successful behaviour, and

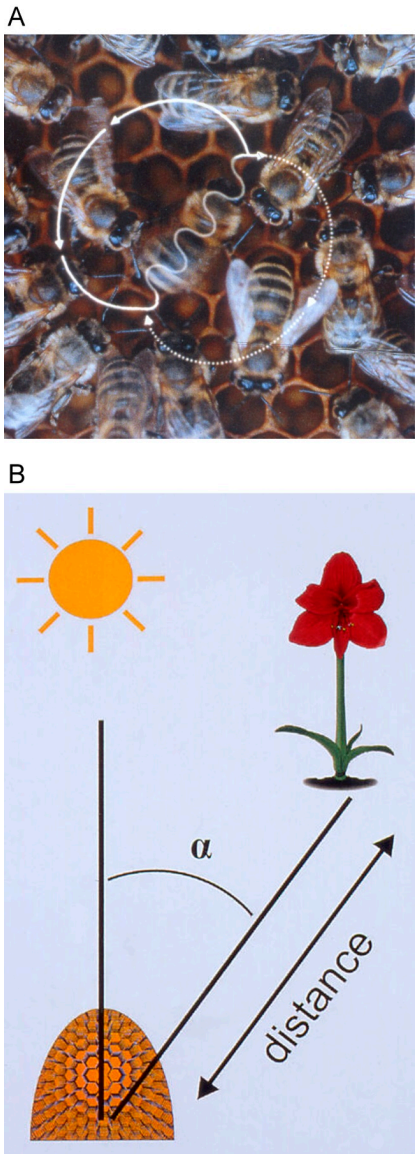

Figure 2.1: How the waggle dance indicates the direction of food. Caption from the original image: "A waggle run oriented 45° to the right of 'up' on the vertical comb (A) indicates a food source 45° to the right of the direction of the sun outside the hive (B). The abdomen of the dancer appears blurred because of the rapid motion from side to side. (Figure design: J. Tautz and M. Kleinhenz, Beegroup Würzburg.)" (Chittka, 2004, p. 898)

their discovery helped win Karl von Frisch the Nobel Prize in 1973 (Karolinska Institutet, 1973). That recruits are capable of finding food at a rate better than chance having followed the relevant dance is an established fact. In what follows, we shall assume that recruits who follow the dance find the indicated food source at a rate better than chance and that the dance evolved for this reason.

2.2.2 Measuring information transmitted by the dance

Soon after the advent of communication theory in the late 1940s, Haldane and Spurway (1954) investigated the informational properties of the waggle dance. I will now describe the aims, methods, and conclusions of that study.

Haldane and Spurway wanted to demonstrate the possibility of deriving informational properties from statistical data. They used honeybee communication as an example because von Frisch had already published the relevant statistics (von Frisch, 1948; von Frisch, 1950, 1952). By 1954, communication theory offered a new perspective on this data from an engineering viewpoint. At the heart of the motivation for the study was the possibility of comparison with other animals. It was hoped that the way information is measured, and the generality of the units of information, might allow for the magnitude of information transmission to be compared between communication systems. Indeed, soon after the study was published Wilson (1962) obtained similar statistical and informational data for fire ants, explicitly comparing his results with those of the earlier work.

Although contemporary work on biological information typically begins by citing Shannon (1948b,c) or Shannon and Weaver (1949), Haldane and Spurway took their leave from Norbert Wiener's *Cybernetics* (Wiener, 1948). Wiener's approach emphasizes intra-system

control over inter-system communication. Since a honeybee colony may fruitfully be construed as a more or less unified entity with unified goals (at least with regard to foraging), it is appropriate to take a cybernetical approach to its behaviour. Haldane and Wiener also happened to know each other personally, which may have contributed to the choice (Dronamraju, 2017, pp. 259-60).

Haldane and Spurway (1954, p. 255) use just one equation to measure information. It is a formula that converts two pieces of statistical data (representing receiver behaviour) into a term for information transmitted by senders. The conversion works as follows. Suppose a group of receivers are about to embark on a foraging run. In the absence of experience they will tend to spread themselves evenly around the foraging domain. Their spatial distribution is represented by a uniform probability distribution (figure 2.2**A**). This is the first piece of statistical data, generated by the modelling assumption that bees without information forage randomly. Note that this is a continuous distribution. The foraging domain has *not* been segmented into discrete 'cells', each with some nonzero probability of being arrived at. Rather, *any* given region – of any size – has a nonzero probability of being visited. Now consider how receiver bees are distributed when they react to a dance. Instead of dispersing randomly throughout the domain, they all fly in roughly the same direction (figure 2.2**B**). This is the second piece of statistical data, generated by the observed distribution of recruits which is assumed to be Gaussian. This too is a continuous function. The 'most popular' direction is the direction of food, but some receiver bees will be a little inaccurate. The conversion equation translates these two distributions into a quantity of information, which can be interpreted as the information provided by senders about the direction of food. It is calculated as follows.

The conversion equation is derived from two formulas taken from Shannon (1948c).

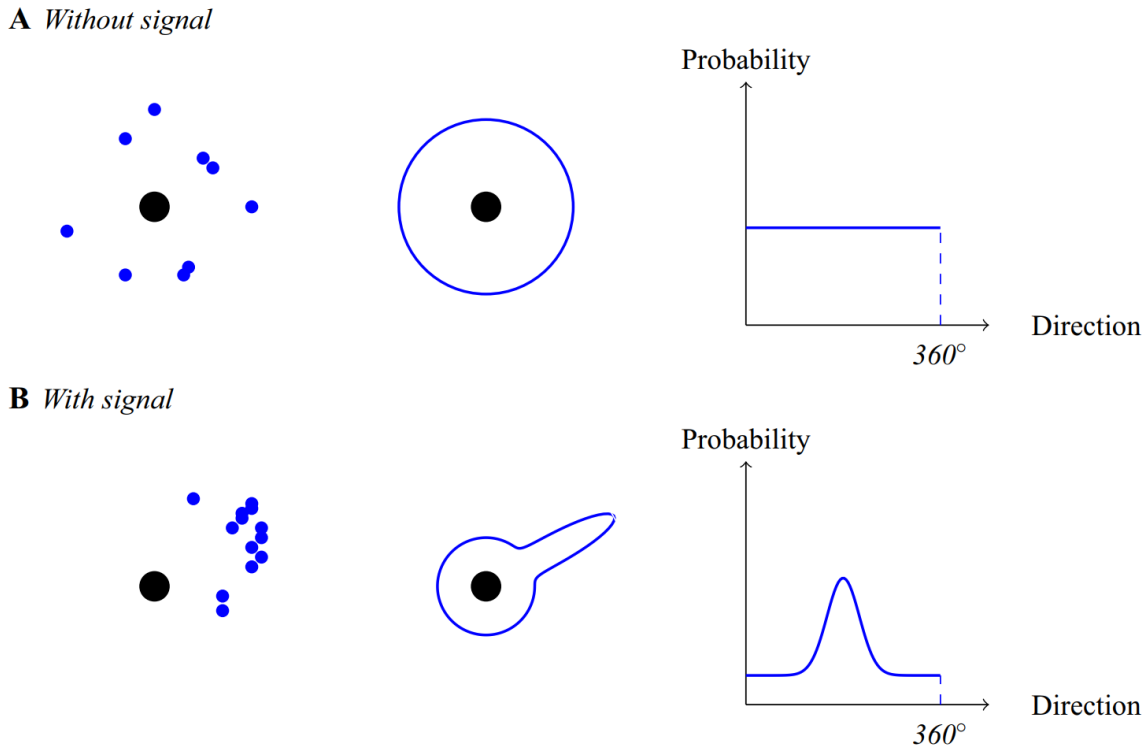

Figure 2.2: How directional information in the waggle dance is measured. **A** Without experience, bees are expected to disperse randomly (small dots) about the hive (black disc). This is represented by a uniform distribution (circle) around the hive. The same distribution in Cartesian coordinates is also presented. **B**. After receiving a signal, receiver bees are much more biased toward the direction indicated by the dance. This situation is represented as a Gaussian distribution. Information transmission is measured by subtracting the entropy of (B) from the entropy of (A) (see equation (2.2) for details). Although individual entropies are quantified relative to the coordinate system (here 360 degrees per circle), transmission rate is not relativised in this way (Shannon, 1948c, p. 631).

These formulas give the entropy of a Gaussian distribution (Shannon, 1948c, p. 630) and the transmission rate of a continuous channel (Shannon, 1948c, p. 637). Haldane and Spurway combine the two into a single equation giving *the transmission rate of a continuous channel when the source entropy is uniform and the conditional entropy is Gaussian*:

$$\text{Transmission rate} = H(X) - H(X|Y) \tag{2.1}$$

$$= \underbrace{\log_2 360^\circ}_{\substack{\text{Prior uncertainty} \\ \text{(figure 2.2\textbf{A})}}} - \underbrace{\log_2 \sqrt{2\pi e}\sigma^\circ}_{\substack{\text{Posterior uncertainty} \\ \text{(figure 2.2\textbf{B})}}} \tag{2.2}$$

The most perspicuous form of this equation is due to Wilson (1962, Appendix, p. 156):

$$\text{Transmission rate} = \log_2 \frac{360^\circ}{\sigma^\circ} - \log_2 \sqrt{2\pi e} \tag{2.3}$$

Here, σ represents the standard deviation of the Gaussian distribution. Informally, it measures the 'spread' of receiver bees about the food source. Granted that receivers can be represented by a Gaussian distribution, σ is the only statistic required to calculate information rate.[1]

An equivalent way of conceiving the situation is to take the uniform distribution as a 'prior' and the Gaussian as a 'posterior'. Before the dance, receiver bees are maximally uncertain about the location of food, hence the prior distribution is uniform. After the dance, the receiver bees' uncertainty has decreased. The magnitude of the decrease in uncertainty is a relation between the Gaussian and uniform distributions.[2]

[1] The appropriate measure for the Gaussian distribution is presented in Shannon and Weaver (1949, p. 89). Haldane and Spurway cite Wiener (1948, p. 62) who provides a general formulation of information in a continuous distribution and does not appear to discuss the Gaussian case explicitly. For a derivation (and explanation) of the equation used by Haldane and Spurway (1954, p. 255) see Wilson (1962, Appendix).

[2] In fact the von Mises distribution would have been more appropriate (Schürch and Ratnieks, 2015). The

By this method, Haldane and Spurway derive a quantity of information about the direction in which food can be found. Substituting the standard deviation of 14.7° (p. 251) into equation (2.3), they conclude that each recruit receives on average 2.5 bits (p. 278) of information about the direction of food. It is worth pointing out that this analysis holds regardless of facts about the cognitive powers of individual bees. On the cybernetic interpretation, information is a measure of how behavioural precision contributes to system goals. Cognitive sophistication can be included or omitted from models of communication, without bearing on the presence of information in signalling systems. The same is true of sender-receiver models in general (Skyrms, 2010, p. 44).

Why is this method justified in the case of honeybees?[3] If a colony is an entity with unified evolutionary goals, it is plausible to think senders obtain a payoff for receiver success. Since receiver success increases when information is transmitted, the dance evolves because of the information it provides. In other words, when 'information' is construed in terms of accuracy, transmission rate is selected for. Nature plausibly selects for accuracy under the circumstances bee colonies have historically found themselves in. On this interpretation, selection has led to increased information transmission.

Furthermore, this line of thought suggests an explanation for increased articulation in the waggle dance over phylogenetically earlier kinds of dance. If articulation supports greater information transmission, and transmission rate is selected for, then articulation can be selected for its informational consequences. Of course, adaptation is always subject to constraints, and the quantity of information transmitted cannot increase indefinitely (Preece and

Gaussian is an acceptable approximation.

[3]Gallistel (2020) offers a more psychologically oriented answer to this question.

Beekman, 2014). Nonetheless, it is justifiably claimed that the benefits of foraging accuracy historically underpinned selection for greater information transmission in the honeybee waggle dance.

In sum, Haldane and Spurway established that informational measures can be derived from statistical data. They justified their use of a measure taken from communication theory by adopting a cybernetic perspective on honeybee colonies. To the extent that a colony's communicative goals are unified, this approach is justified. Since 1954 more work in a similar vein has appeared (Beekman et al., 2015; Riley et al., 2005; Schürch and Ratnieks, 2015). Regardless of differing results, what matters here is whether Haldane and Spurway's model was a good one and whether their interpretation of information was valid. I have argued that the answer to both of these questions is yes. The next section responds to two objections.

2.3 Initial objections to Haldane and Spurway

Skepticism of the validity of these results continues (Pfeifer, 2006; Sarkar, 2013). Even optimistic scholars, as well as those on the fence, cite Pfeifer (2006) as having presented technical challenges to the use of information theory in biology (Reading, 2011, p.149 fn.6) (Stegmann, 2013b, p. 143) (Wiley, 2013a, p. 118). I aim to show that although there surely will be technical and methodological hurdles, they are not the ones presented by these objections.

2.3.1 Objection 1: Arbitrary Coordinate Systems

In this section I show how the bee dance is both iconic and continuous and how signals of this kind contain measurable information. This undercuts one skeptical argument against the use of information theory for animal signals in general, and the bee dance in particular, namely that the models we employ to quantify signal information are somehow arbitrary. Discrete models appear to be arbitrary, but continuous models do not suffer from the same defect.

The skeptical argument due to Pfeifer (2006) and repeated by Sarkar (2013, §7.3) runs as follows. Honeybee signals indicate, among other things, the direction of a food source relative to the hive. The quantity of information about direction in a signal depends on how precise the signal is. When modeling bee signals, therefore, the amount of information we will obtain depends on how finely we carve up 'direction space.' But we have no a priori guidance as to how finely to carve up the space. Different divisions will give different measures of information, but none will have priority. Sarkar suggests we could test for individual bees' perceptual acuity, and divide the space into segments of a size roughly discriminable by the bees themselves. He claims Haldane and Spurway did not do this, and that even if they had, the placement of these individual direction-segments would have been arbitrary. By shifting them a half-step clockwise, we obtain a totally new division of direction space, which will transform our statistical data into different informational measurements. Since no carving of the space is privileged, there can be no privileged conversion of statistical data into informational quantities. As a result, information measurements are to some extent arbitrary, and cannot be relevant for selective explanation.

To make the objection clearer, consider one of Pfeifer's models (Pfeifer, 2006, p. 325).

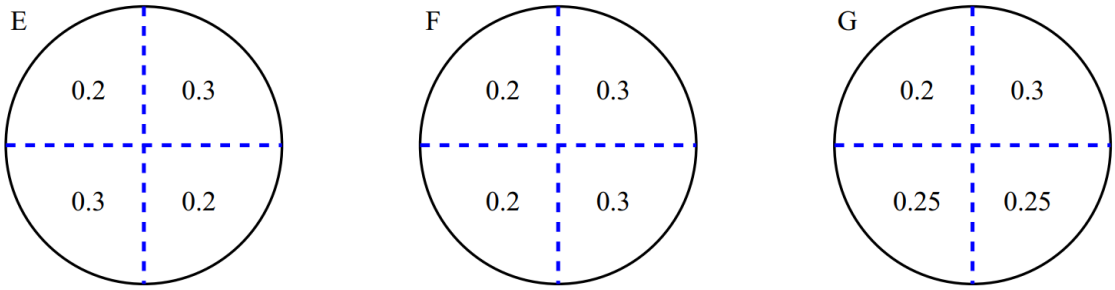

Figure 2.3: Signal probabilities for model **EFG**.

Suppose there are three types of bee, **E**, **F** and **G**.[4] They each need to signal the location of food, and each signal indicates a direction around a circle centered on the nest. Further, they differ in the probabilities of signaling in each direction, as per figure 2.3.

Described this way, each type transmits a certain amount of information on average:

- **E**: 1.971 bits/signal
- **F**: 1.971 bits/signal
- **G**: 1.985 bits/signal

We see that **G** transmits more than **E** and **F** on average. However, if response behavior is distributed uniformly about the segments, we can individuate signals by cardinal points instead, as per figure 2.4.

Described this alternative way, the information measurements are:

- **E**: 2 bits/signal
- **F**: 1.985 bits/signal

[4]The original example mentioned ants. Here and below I substitute bees without loss of generality. I retain the alphabetical labeling of types for ease of comparison with Pfeifer's article. Model **ABCD** is not relevant to the argument and has been omitted.

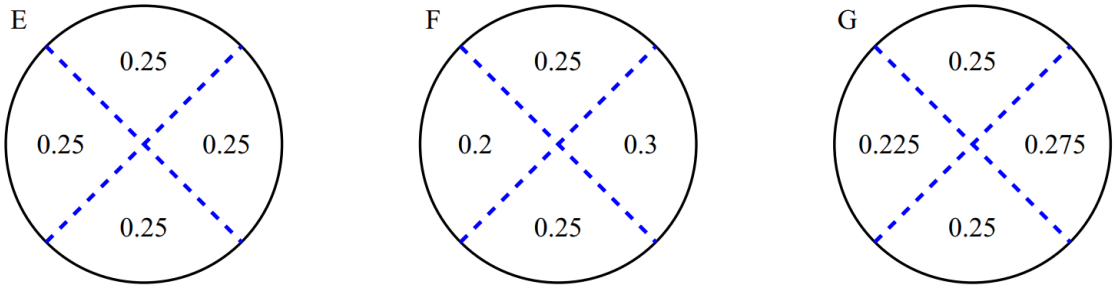

Figure 2.4: Signal probabilities for model **EFG**, alternative description.

- **G**: 1.996 bits/signal[5]

Notice that the order of greater/lesser information transmission has changed. Now **E** transmits more than **G** types, followed by **F** types last. The subsequent argument is fairly obvious. If signal individuation is down to the modeler's discretion, and information measures are sensitive to individuation, then information transmission is model-relative.

Though there are many ways for depictions of the world to be relativized to the model used to express them, and not all render hypotheses inconsequential, the kind of relativity implied by this objection is problematic. A selective hypothesis is supposed to pick out a property and state why it persists in the lineage. In this case, the property is the quantity of information contained within signals. If that quantity cannot be said to be greater or lesser except relative to the model chosen to represent the signal, we need independent grounds on which to choose one model over another. Pfeifer claims we have no such grounds since no carving of the space is privileged: "if different ways of carving signals result in different orderings of the measurements, then it is unclear how the measure could be used in explaining

[5] A typo in the published version of Pfeifer (2006, p. 325) erroneously cites this value as 1.96 bits/signal.

how information transmission is selected for" (Pfeifer, 2006, p. 325, col.1). As a result, no selective hypothesis adverting to information transmission is recommended over any other.

Response to Objection 1

There are two ways to read the objection. I will argue it fails in either case. On the first reading, it turns out recarving the space does not alter the informational measurement. The objection fails on mathematical grounds. On the second reading, recarving the space alters the measurement as required, but some carvings will more accurately match the facts than others. Selective hypotheses adverting to informational measurements are then verifiable, establishing the required link between model and reality. The objection fails on both readings. Moreover, the dilemma is not accidental. It is a reflection of Pfeifer's assumption that signals are discrete and indicate direction symbolically. In fact, signals in this system are continuous and indicate direction iconically.

Consider the first reading. The objection rests on statistical differences between the two carvings. However, the second partitioning does not seem to have the entropy assumed by the second informational calculation. If the distribution of insects across each partition in the first carving is uniform, the distributions in the second will *not* be uniform. To see this, consider figure 2.5**A**. Blue solid lines represent insects observed in each segment as a proportion of total insects observed. These densities would not change even after the partitions are shifted, so the new distributions are nonuniform lines with a step-change in the middle of each partition. The information rate derived from such stepped distributions is the same as that calculated from the original partitioning. As a result, the proposed calculation is incorrect and does not have the consequence that information measurements are model-

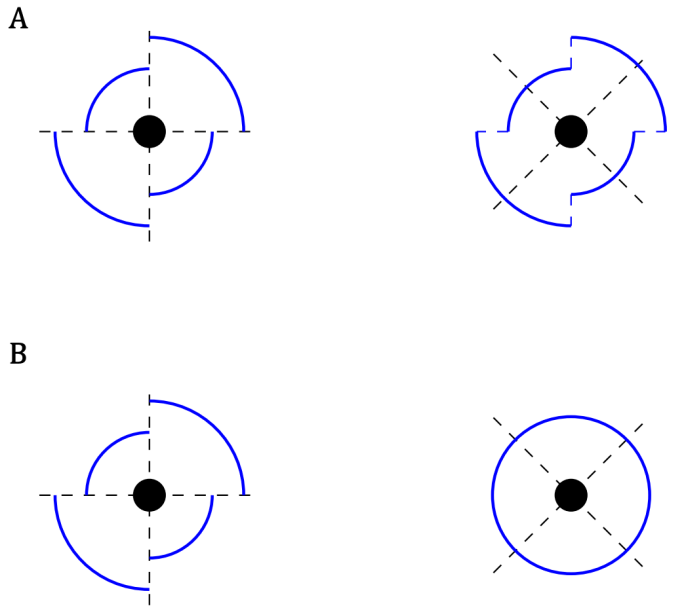

Figure 2.5: Two ways to read objection 1. In either case, the objection fails. Blue solid lines represent insects observed in each segment as a proportion of total insects observed. **A**. The different carvings of the space entail the same informational results (because the blue lines do not change). The objection is then incorrect on mathematical grounds. **B**. The different carvings of the space entail different informational results (because the blue lines are different). But these different carvings now entail different falsifiable assumptions. The objection fails because the models' assumptions are verifiable, not arbitrary in the manner required for the objection to go through.

relative.

There is another way to read the objection, however. Suppose the scientists were faced with a choice of partition to use while conducting the experiment. They can set up their four detectors in either of the two configurations depicted by model **EFG**. Due to restrictions in the precision of the detectors, they are forced to assume a uniform distribution within each segment in either case (figure 2.5**B**). Given these conditions, it is true that a scientist who opts for the first configuration would derive different informational measurements than

one who chooses the second. Different selective hypotheses result, but one hypothesis will be better. One carving entails more accurate modeling assumptions, since the regions it describes in fact have a more uniform distribution of insects passing through. Alternatively, both carvings could be equally wrong. This is importantly different from being arbitrary, since there will be some other carving that is better than both. In any event, a more precise detection method would capture more fine-grained insect movements, giving rise to more accurate informational calculations.

The objection fails on both readings, a dilemma that brings out a key difference between model **EFG** and the model used by Haldane and Spurway. The latter employed a discrete approximation of a continuous model, using an equation of continuous rather than discrete entropy. Although the statistic they plugged into the equation was derived from an experiment using discrete partitions, it is a different way of modeling the situation than that typified by **EFG**. Discrete approximations to continuous models become more accurate as they become more fine-grained (Schürch and Ratnieks, 2015, Fig. 1, p. 3). If the insects of model **EFG** were observed again using detectors capable of distinguishing 8, 16, or 32 partitions, the resulting information measurement would become more precise. It would not increase indefinitely, as per Pfeifer's approach.

To reiterate, Pfeifer introduces a situation in which signals are discrete and represent directions symbolically. Her objection is then that we cannot possibly come to know how symbols and directions correspond, and any choice destroys the link between model and reality. In contrast, the biologists make the reasonable and widely accepted assumption that bee signals are continuous and represent directions iconically. Once a continuous model is employed to represent a signaling system, discrete partitions can be overlaid to retrieve data.

The more fine-grained these partitions, the more accurately the data captures the continuous model lying underneath.

A different but equally mistaken conclusion might be drawn from the equation used to convert statistical into informational data. To measure the accuracy of the waggle dance, Haldane and Spurway took a circle centered on the nest and measured the proportion of that circle covered by insects that had received a signal (recall figure 2.2 and equation 2.3). A smaller circle proportion covered by food-seeking receivers means more accurate communication, hence a greater amount of information transferred. Pfeifer objects to the appearance of the number 360 in equation (2.3):

> It is assumed in both cases that there are 360 possible messages about direction corresponding to the 360 degrees surrounding the nests of the ants or bees [but] it is not clear why they should be divided into 360 different possible signals, as opposed to 180, 720, or some other number.
>
> Pfeifer (2006, p. 342)

It is easy to see the problem if the objection hits its mark. The division of circles into 360 equal segments is a human convention. If it is used as the basis of an information calculation, any derived quantities must be arbitrary as well. Conventional measurements cannot enter into explanations that pertain to natural selection, since nature is blind to our conventions. Consequently the use of the number 360 in equation (2.3) is inappropriate. Again, citing Pfeifer, Sarkar reiterates the objection (Sarkar, 2013, p. 196).

However, the interpretation of equation (2.3) that underpins this argument is incorrect. The number 360 is employed only as a convenience for the reader. The equations in which it appears are designed to calculate the proportion of the circle covered by outgoing insects. Proportion is blind to the units used to calculate it since dimensions cancel in the division.

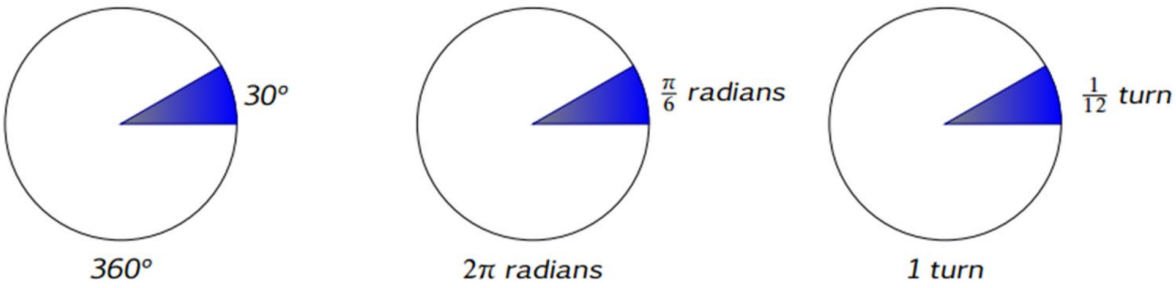

Figure 2.6: The proportion of a circle covered by receiver insects can be calculated using any units, whether degrees, radians, or as a fraction of a turn. In equation 2.4, σ (Greek letter sigma) represents the angle of the shaded segment. Hence the first logarithm in equation 2.3 represents the length of an arc as a proportion of the circumference of the circle, which is not relative to the conventional measure of 360 degrees.

Indeed, Haldane and Spurway first represent the full angle in radians as 2π before converting it to 360 degrees for expository clarity (Haldane and Spurway, 1954, p. 255). Unfortunately they use the same symbol, σ, in both cases, making it less than obvious that the units have changed from radians to degrees. Equation (2.4) and figure 2.6 demonstrate the equivalence of these different measures.

$$\text{Circle proportion} = \frac{\sigma_{\text{degrees}}}{360} = \frac{\sigma_{\text{radians}}}{2\pi} = \frac{\sigma_{\text{turn}}}{1} \tag{2.4}$$

As a result the use of the number 360 does not indicate an assumption of 360 messages. It is true, however, that the *precision* of the calculation increases as the precision of σ increases. Pfeifer raises a related complaint that the statistical data available to Haldane and Spurway are given in 15 degree increments, which is too coarse-grained to provide a reliable measure. However, the biologists take this into account, and the error margins they give are consistent with their estimations. As per Schürch and Ratnieks (2015, p. 3), more precise measurements

would lead to more accurate calculations. In sum, this objection is not sustained.

Overall then, the accusation of arbitrariness toward coordinate systems used to measure the directional component of the bee dance cannot be upheld. Continuous models are available for use where appropriate. Haldane and Spurway employed such a model. Though the literature on sender-receiver models emphasizes discrete signal sets, this is not a necessary feature of the framework. Plausibly, many signals in nature can be profitably investigated with continuous models. Discrete models are often best construed as approximations to the underlying continuous representation. In the next section I deal with a second major objection to the use of information theory in the study of biological communication.

2.3.2 Objection 2: Precision and Accuracy

In this section I show how quantifying information in signals is not just a case of quantifying the precision of receiver behavior. One route to skepticism of the utility of information theory in biology is based on the claim that behavioral precision does not entail accuracy, hence precision alone entails nothing about fitness. I show that information measurements, far from being solely a measure of precision, are a quantification of accuracy. Accuracy, which in this case translates into foraging efficiency, is clearly relevant for fitness.

The objection, again due to Pfeifer, runs as follows. The biologists used statistical data in their calculations, and the statistic they employed is the spatial distribution of signal receivers. A greater proportion of receivers clustered closer to the target entails a greater quantity of information transmitted. By measuring the difference between this more 'focused' distribution, and the otherwise random distribution of receivers around the circle, the biologists quantify information. Pfeifer complains that mere increased precision of receiver behavior

might not be selectively relevant. Suppose bees clustered closely together in a region that lacked food. Their behavior would be focused in a statistical sense, yet would not contribute to fitness. Behavior can be highly specific yet selectively neutral, or even detrimental. To support the objection Pfeifer presents model **HIJ**.

Suppose there are three kinds of bee, **H**, **I** and **J**, each of which can signal to the east and west and do so with equal probability. Suppose **H** types always go east when signaled east and west when signaled west, but **I** types respond improperly by going west when signaled east and east when signaled west. In addition, **J** types go east with probability 0.7 when signaled east, otherwise west, and vice versa. The problem highlighted by this model is that **H** and **I** types transmit the same amount of information, namely 1 bit/signal, whereas **J** types transmit much less, around 0.119 bits/signal. Despite this, we would expect **J** types to perform better than **I** types because they successfully obtain food more often on average. In consequence, it is unclear how information transmission can have relevance for fitness.

2.3.3 Response to Objection 2

We can respond to this objection by noting that the increased focus of the posterior distribution is centered on a target that has selective relevance, such as food. Wilson (1962) confirms that this assumption is encapsulated by the model. The parameter used to measure information, σ_d, is directly relevant for accuracy:

> Note that σ_d refers to dispersion of following workers with reference to the actual target and hence is a direct measure of the accuracy (content) of information transmission. The honeybee data of von Frisch and his co-workers are of this nature.
>
> Wilson (1962, Appendix, p. 157, col. 1)

Insofar as receiver bees have a greater probability of finding food after having watched a dance, information is transmitted by dancing bees. It is this notion of transmission that constitutes grounds for application of the information concept, and related mathematical tools, in studies of biological signaling.

What about model **HIJ**? It seems to show that biologists might inadvertently attribute a high rate of information transmission to a colony that fails to put its signals to good use. As a claim about methodology, this is implausible. If it is assumed that signals used by type **I** are indeed food signals, it would be mysterious why receivers behave so aberrantly. Where the colony's goal is finding food, signals generally increase the probability of doing so. When they do not, they are selected against. **I**-type signals could hardly be regarded as signals about food since the probability of finding food is not raised when they are sent.

Like any piece of behavior, a signal cannot be selected for if it does not contribute to fitness. By hypothesis **I**-type signals do not contribute to fitness, therefore they cannot be selected for. As a result they would not contribute to a behavioral ecologist's selective hypothesis; they would be simply anomalous.

To reiterate, the biologists' models embody a crucial assumption: that the target has selective relevance. In this section, two mathematical objections have been overturned. I now turn to broader considerations about the use of informational concepts in animal communication theory.

2.4 Are information and meaning appropriate concepts for animal signals?

The most pressing challenge to the realist about biological information is the lack of a clear definition of that concept in behavioural ecology. A little background is required to motivate this objection.

2.4.1 Scepticism about information in behavioural ecology

By the late 1970s, animal signals were typically defined in terms of information. Signal evolution was thought to mainly proceed via ritualisation of cues. Since ritualisation is a process of coadaptation, signals were viewed as primarily cooperative, despite the possibility of deception and exploitation. In the face of this orthodoxy, Dawkins and Krebs (1978) offered a different approach. They proposed that communication be defined as one animal controlling another, typically by exploiting perceptual mechanisms designed for other purposes. Due to the individualistic operation of natural selection, cooperation and coadaptation are rare (Dawkins and Krebs, 1978, p. 289). Consequently, we should adopt an individualistic account of communication, one that places the sender front and centre.

More recently, in a series of individual and joint papers, Michael J. Owren, Drew Rendall, and Michael J. Ryan (hereafter ORR) present significant challenges to the use of informational concepts in animal communication studies (Owren et al., 2010; Rendall and Owren, 2013; Rendall et al., 2009; Ryan, 2013). Information, they claim, is an insubstantial metaphor that cannot do explanatory work and often misleads us as to the nature of signalling behaviour. Several behavioural ecology texts go so far as to define communication in terms

of information, without ever fully explicating the latter (see Rendall et al. (2009, Table 1) for examples). Instead, ORR propose a definition of signalling akin to that of Dawkins and Krebs, in terms of the influence one animal exerts over another.[6] Further scepticism is advanced by Sarkar (2013), who discusses the honeybee waggle dance along with wider issues of animal signalling. Sarkar claims different informational concepts have been conflated in the literature, and that as a result informational measurements do not capture the quantities claimed by those who employ them.

ORR point out that many authors use the quasi-technical term 'information' in an unconstrained manner (Rendall et al., 2009, Table 2) (Rendall and Owren, 2013, Table 6.1). As a result those authors often switch between different concepts masquerading under the same label, or gesture at unsubstantiated explanations. In particular, authors often invoke 'information' in a technical sense that quantifies correlations between events. This, ORR claim, is Shannon information, which cannot be central to definitions of signalling because it can be used to describe any correlated events, biological or otherwise. Sarkar (2013) too distinguishes Shannon information (he calls it MTC-information for "mathematical theory of communication") from semantic information, and repeats ORR's contention that theorists often conflate the two. The distinction has also been flagged as between "syntactic" and "semantic" information, the former being agnostic about the meaning of signals or symbols whose transmission it quantifies (Morton and Coss, 2013, p. 211 & p. 229). Since invocations of information are inconsistent, and nobody has provided a resolute definition, it seems

[6] Stegmann's edited volume (Stegmann, 2013a) characterises the debate as between information-based and influence-based definitions. However, as pointed out by several entries in that volume, information and influence do not form a strict dichotomy. Here I am primarily concerned with establishing grounds for the attribution of information, so I make little mention of influence-based definitions.

better to dispense with the notion.

A similar fate befalls the concept of encoding. Animal signals are sometimes said to carry information in a form that a receiver must decode in order to obtain. But it is rarely made explicit what is meant by an *encoded message* in a biological setting. By failing to specify what is being invoked, theorists draw on a "vague, elastic and insubstantial" concept (Owren et al., 2010, p. 758) that cannot do the explanatory work required. As with "information," then, talk of "encoding" does more harm than good. Both can be discarded without losing explanatory power. Rendall and Owren (2013, pp. 171–2) make an even stronger claim: "Ultimately, notions such as information and coding cannot be cashed out in terms of standard concepts used in biological and evolutionary theory."

2.4.2 Responding to information scepticism

One thing worth mentioning here is that ORR's wider aim, to critique anthropocentrism and linguistic metaphors in animal signalling theory, is laudable. Animal communication should indeed be understood on its own terms, and should not be treated as a pale imitation of human natural language. It is interesting, however, that ORR consider the use of information theory to fall into this anthropocentric paradigm, since several theorists present the story precisely the other way round. The "conduit metaphor" that ORR discuss is a term introduced by Reddy (1979) in part to highlight the limitations of modelling human communication as a signalling channel. Reznikova (2017) introduces information-theoretic methods in studies of communication between ants in an attempt to get away from language-inspired metaphors and methodologies. Harms (2004) promotes formal methods for capturing meaning without recourse to translation, since translating animal signals into human language can

only mislead us about the nature of meaning. All these theorists note the disparity between communication theory and human language. In chapter 7 I continue work in that tradition, arguing that animal communication can be described by the relevant mathematics precisely because information theory captures something far more general than human engineering constructs. It should become clear that information theory, rightly interpreted, is no more anthropocentric than any other branch of mathematics. Furthermore, in chapter 3 I argue that claims about the irrelevance of information measures for "semantic information" are deeply mistaken.

For now, we can show once more that both information and encoding are clearly defined in the honeybee case. As discussed at length in section 2.2, the measurement of transmission rate captures the accuracy of receiver bees, hence the efficiency of foraging behaviour. Haldane and Spurway's insight was to define an entropy over the space of relevant behavioural outcomes such that when entropy is reduced by a communicative act, the measure of information rate is simply a measure of successful functional performance. This is an easy measurement to perform in the honeybee case because foraging efficiency is linked to spatial accuracy, and spatial accuracy is comparatively easy to ascertain. So information can be defined, at least in this case, in terms of functional performance, rendering it both tractable and relevant for selection.

Consider now the concept of encoding. After a period of controversy in the middle of the 20th century, it became widely accepted that the different components of the bee dance correspond to different spatial relations between the hive and the indicated food source. These aspects of the dance vary with those spatial relations in a principled way. If they did not, receivers would be unable to reliably exploit the relevant food source. Although there

are several other factors affecting transmission rate, the mapping between spatial relations and dance features clearly affects it too.

Could the waggle dance mapping be an encoding in the relevant sense? In section 5.5 I argue that mapping rules are not strictly encodings, so we need to be careful here. There are cases where encodings and mapping rules coincide. The central model is an example. Codes in the central model are syntactic transformations from source string to codestring. The mapping rule is a semantic transformation that also takes source strings, considered as world affairs, to codestrings, considered as signals. The sense of encoding applicable to the waggle dance has a similar character. Scouts experience the location of food and dance accordingly. There is both a direct causal link between food location and dance (the 'encoding'), and a functional mapping between them that explains receiver success (the mapping rules). It is conceivable that, beyond the central model, the term 'encoding' more naturally applies to the mapping rules for the signalling system. Nonetheless, it is important to distinguish *how a signal is produced* and *what it is supposed to bear a mapping relation to*.

Different mappings can play the same functional role. Consider the round dance, an alternative to the waggle dance used when food sources are very close. This dance has no directional component, meaning that on average its transmission rate would be lower than the waggle dance. Greater articulation affords greater transmission rate, which presumably contributes to the explanation why the waggle dance is articulated into different components. In general, when behavioural ecologists talk of information being encoded in animal signals, they are referring to a mapping between signal and world. The mapping explains how receiver behaviour covaries aptly with the state of the world.

Owren, Rendall and Ryan offer further considerations on the role of information in ani-

mal communication studies. One of the strongest arguments launched by information sceptics is the problem of conflicting interests. Many animal interactions that behavioural ecology counts as signalling occur between organisms whose evolutionary interests are imperfectly aligned. The informational perspective was licensed for the waggle dance in part because bees belonging to the same colony can be treated as having unified interests in the context of foraging. When the assumption of cooperation is relaxed, what becomes of information? I discuss this question in chapter 7, especially section 7.2. To anticipate, the informational perspective is most relevant for paradigmatic signals. But the fact that some – perhaps many – signals in nature are sent between organisms with conflicting interests does not jeopardise the salience of information in the cooperative case.

2.5 Conclusion

There is at least one case in which information is literally transmitted by an animal signal. The waggle dance improves the performance of watching scouts, and evolved for this reason. Information in the dance is measured in terms of the amount of functional improvement it brings about. Because functional improvement is defined in terms of receiver bees finding a particular source of food, its extent can be measured and converted into informational units. Scepticism about the model underpinning these calculations is sometimes mistaken and therefore unwarranted. However, the strongest challenge to universal application of the information concept in behavioural ecology is the threat of divergent interests between signallers. This problem has not been dealt with yet, and will be addressed in chapters 6 and 7.

Chapter 3

Information and content

3.1 Introduction

Philosophical approaches to content with an eye on the science of information typically begin with something like this thought. Information as defined mathematically has a role to play in scientific theorising, but it does not possess the requisite properties to be identified with content. Information might form part of a definition of content, but information alone is not content. In particular, mathematical information says nothing about correctness conditions (Godfrey-Smith and Sterelny, 2016; Lean, 2014; Piccinini and Scarantino, 2011) (Shea, 2018, p. 13). The strongest version of this idea is the **irrelevance claim** – the claim that mathematical measures of information are irrelevant for philosophical theories of semantic content.

In reaction to this sceptical view, others have argued that informational properties just are semantic properties (Isaac, 2018; Skyrms, 2010). Their claim is supported by the association of statistical vectors with signs, that supposedly capture their semantic content. These vectors, which I call **statistical content** and describe in more detail in appendix C, capture correlational relationships between a sign and world affairs. By associating lists of correlated states with a sign, Isaac claims we capture the sign's semantic content, what it 'says about the world' (Isaac, 2018, p. 2). On this view, semantic content is universal: virtually every event carries it in some form or other, because virtually every event is statistically correlated

to at least some degree with others.

In this chapter I challenge both the sceptical view and the universality response. Rather than simply identifying information with content, I will redraw some traditional boundaries between concepts. Many of the properties claimed to be important for content can indeed be found in the mathematical definition of information. But while the association of signals with **correctness conditions** is ineliminable from communication theory, the same is not true of all applications of informational measurements. Statistical content does not play the explanatory role required of semantic content, because statistical content does not specify correctness conditions.

The chapter proceeds as follows. I argue that communication theory can be treated as a branch of game theory (section 3.2). Agential behaviours – *strategies* in the language of game theory – appear in communications engineering as *codes*. In section 3.3 I introduce, and explore the provenance of, two popular sceptical claims about the relevance of information theory for philosophical accounts of semantic content. Then in section 3.4 I argue against both sceptical claims, highlighting several interpretive mistakes they are predicated upon. In section 3.5 I push against the prevailing view that distinguishes mathematical and semantic concepts of information. Finally, section 3.6 responds to objections. While this chapter is broadly negative – arguing against scepticism and universalism – chapter 4 offers a positive view of teleosemantic mapping relations and their explanatory role.

3.2 Communication theory as game theory

Game theory was originally conceived as normative. von Neumann and Morgenstern (1944) published the foundational text, since when game theory has been applied in economics. In this domain, game theory provides advice for **rational agents**, strategic interactors who represent utilities in the form of preferences and combine them with beliefs to choose optimal actions. Although it can be interpreted as predictive, describing what humans faced with economic decisions will do, its highly idealised models are rarely accurate as representations of real agents. Instead, it can be treated as a way of determining the best course of action in a given situation.

Game theory's descriptive features come out most clearly in its counterpart, evolutionary game theory. Instead of rational agents, the denizens of evolutionary models are strategic interactors produced by selection. **Evolutionary agents** do not have preferences in the sense of cognitively represented utilities. Their behaviour can nonetheless be explained in terms of the payoff schedule that produced them. These payoffs play the role of utilities, and are commonly associated with evolutionary fitness.

3.2.1 Codes are pairs of strategies

In evolutionary game theory, selection gives rise to strategies. Signalling strategies are pairs of behaviours that serve to coordinate sender and receiver activity. The central model of communication theory is a sender-receiver model. Signallers and their strategies are chosen by rational agents – human engineers. This is no less an instance of selection than evolution by differential reproduction. Communications engineers do not describe senders and receivers as performing proper functions. Nevertheless, in both cases design is an ineliminable part of

the analysis. It may be tacit, it may be naturally selected, but design is still design.

The central model falls into the sender-receiver paradigm, but it underspecifies how exactly the bare framework is to be overlaid. For example, consider the simple correspondence depicted in figure 3.1B. This is the canonical interpretation of the central model in terms of game theory. Martínez (2019) uses this representation to argue for links between game theory and communication theory. Earlier, Godfrey-Smith suggested such a connection, citing Lewis (1969) as the originator of sender-receiver games:

> Shannon took for granted the sender and receiver roles, and gave a theory of the properties of channels that could achieve coordination between them; Lewis took for granted the possibility of a channel, and gave a first account of how agents could come to play the sender and receiver roles – how these roles could be stably occupied.
>
> Godfrey-Smith (2011, pp. 1289–90)

Although alternative representations are available (some of which are depicted in figure 3.1C and D), in what follows we shall be interested in the canonical interpretation.

Consider the behaviour of the encoder, as a sender, and decoder, as a receiver. A shared code is a coordinated pair of strategies. Strictly, an encoding is a function from a string in the source lexicon to a string in the code lexicon. Conversely, decoding is a function from a string in the code lexicon to a string in the source lexicon. Depending on the channel and the required accuracy, decoding might not precisely 'undo' encoding. The result might not be exactly the same string, but it might be close enough.[1] But the decoding function is chosen together with the encoding function in order to solve a joint problem. Desirable properties

[1] See chapter 6 for details of rate-distortion theory, the branch of communication theory that defines what 'close enough' means and shows how to achieve it.

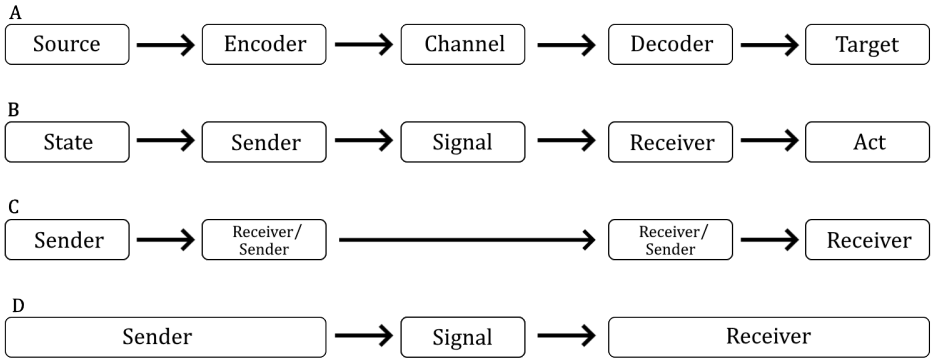

Figure 3.1: Three different ways to map the central model (depicted in **A** with noise node omitted) onto the sender-receiver framework. **B** shows the canonical representation. In **C**, each pair of causal nodes in the central model is treated as a sender-receiver system; each arrow is therefore a signal. In **D**, the source and encoder are collapsed to a single sender node and the receiver is correspondingly amalgamated; there is therefore no notion of encoding and the state and act are unspecified.

of signals, such as unique decoding, generate constraints on optimal coding methods. All of this appears in introductory texts on communication theory (Cover and Thomas, 2006, §§5, 10) (MacKay, 2003, §§5-18, Part VI). These are problems of joint proper function, and they are solved by jointly selecting sender and receiver behaviour.

An immediate worry is how to interpret utility (or payoffs) in a communication theory context. This highlights our key difference between communication- and game-theoretic analysis. Game theory equivocates between 'what *will* the players do?' (the evolutionary or descriptive approach) and 'what *should* the players do?' (the rational or normative approach). Communication theory asks the second of these. It is a prescriptive theory of engineering design. As a result, there must be assumed payoffs, even if they remain tacit. And indeed there are, lurking in the background of the original statement (Shannon, 1948a) but later brought to the fore:

> Consider a situation in which there is a measure of the fidelity of transmission or the "distortion" between the original and final words. [...] The [distortion] quantity d_{ij} may be thought of as a "cost" if letter i is reproduced as letter j.
>
> Shannon (1959, p. 326)

To anticipate somewhat: the goals of engineers, specified in terms of distortion, imbue signals with content. This is *in addition to* any content the original symbol string may happen to have. We will return to these costs in chapter 6, when we consider what happens when costs diverge for sender and receiver.

I want now to show that codes are a special case of strategies. In signalling games, a strategy is a function from a world state to a signal or from a signal to an act.. By casting the source string as a world state, the code string as a signal, and the decoded string as an act (figure 3.1B), the result follows immediately.

Do source strings really count as world states? As far as I can tell, there really is no difference between the formal properties of world states in signalling games and the formal properties of source strings in the central model. All that is required is a finite distribution over possible states, which we identify with the probability distribution over possible source strings. It just so happens that the world state is made up of concatenations of symbols rather than, say, the occurrence of predators. Martínez (2019) makes much the same point.

There is a further sense in which codes are a special kind of strategy. In addition to dealing with strings rather than world states, they are selected *together* rather than being produced by distinct selection processes acting on each agent individually. Codes are encoding-decoding *pairs*. Typically this assumption is tacit. For example, Cover and Thomas (2006, p. 183) define communication as occurring when "the physical acts of A have induced a de-

sired physical state in B." Game theorists (not to mention behavioural ecologists) would be forgiven for interpreting this to mean that the changes in B were desired by A. But there is no distinction between what A and B want in this context. A human engineer is the sole 'source' of normativity for the system, designing both agents to achieve the same end. We cannot distinguish between the interests of A and B.

3.3 Origins of scepticism

In this section I identify two popular sceptical claims about the application of informational measures in the functional sciences, and trace them back to two true statements about communication theory and information theory. In the following section I argue that both sceptical claims are false. They are not entailed by either of the two true statements or their conjunction.

3.3.1 Two sceptical claims

The sceptical claims I take issue with are:

> **Irrelevance**: Formal tools from communication theory *have no relevance for* correctness conditions.
> **Insufficiency**: Formal tools from communication theory *are not sufficient* to specify correctness conditions.

Both claims appears to be prompted by two propositions which are indeed true. For convenience I label them as follows:

> SHANNON'S WARNING: In the central model, the meanings of source strings are irrelevant to the engineering problem of communication.
> UBIQUITOUS INFORMATION: Formal measures from information theory (such

> as surprisal, entropy and mutual information) can be applied to any statistical variable (not just signals).

I will demonstrate that contemporary advocates of the sceptical claims are drawing directly on the two true propositions. Through a history of mutation and misunderstanding, the sceptical claims have come to seem foundational, when they are anything but. Philosophy has made a collective deductive misstep: neither of the two true propositions, nor their conjunction, entail Irrelevance or Insufficiency.

3.3.2 Shannon's Warning

Shannon published his foundational texts in 1948 (Shannon, 1948b,c). In the first introduction he writes:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another 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 (1948b, p. 379), emphasis original

Clearly *message* in this context refers to a source string. Shannon warns that the semantic properties of lexical elements do not affect the process of transmitting and reconstructing them.

In 1949 Shannon's papers were released in a single volume with prefatory remarks by Warren Weaver (Shannon and Weaver, 1949). One of Weaver's comments expands on Shannon's earlier technical statement:

> In fact, two messages, one of which is heavily loaded with meaning and the other of which is pure nonsense, can be exactly equivalent, from the present

> viewpoint, as regards information. It is this, undoubtedly, that Shannon means when he says that "the semantic aspects of communication are irrelevant to the engineering aspects." But this does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects.
>
> Shannon and Weaver (1949, p. 8)

There are a couple of things worth noticing here. First, Weaver slightly misquotes Shannon ("*the* semantic aspects" instead of "*these* semantic aspects"). In context the mistake is insignificant, because the preceding sentences demonstrate that Weaver interprets the claim accurately. Out of context, however, the misquotation can be read as stating categorically what I deny: that *no semantic properties are relevant* to the well-functioning of engineered communication systems.

Second, the final sentence of the quote is rather mysterious. It seems to suggest that there is *some* connection between the engineering problem and semantic issues. The only philosopher I am aware of trying to interpret this part of the quote is Dretske (1981, p. 41) (see below).

Shannon's Warning is salient in a system with two lexicons. The central model is an example. Symbols of its source lexicon might be able to be combined in order to produce messages that are meaningful in some sense, but any such meanings are not represented in the mathematics. Those meanings therefore do not play a part in the statement or solution of the fundamental problem of communication.

One might think that the meanings of source strings *could* play a role in solving this problem. An intelligent observer receiving the noisy signal SHALL I COMPARW TGEE TO A SUMNERS DAY might be able to reconstruct the original on the basis of its presumed meaning. The point of Shannon's Warning is not to rule this out, but to circumscribe the

statistical aspects of the problem.[2]

The ramifications of Shannon's results were understood immediately. In London, a series of symposia were established to discuss the theory and its implications. As Donald MacKay later noted, these early meetings were one source of what I am calling the Irrelevance claim:

> By the time of our Third London Symposium on Information Theory in 1955, it had become something of an accepted saying that 'information theory has nothing to do with meaning'. The time seemed ripe to question this hardening dogma...
>
> MacKay (1969, p. 79)

Despite MacKay's intentions, the dogma hardened rapidly. Recently, Freeman Dyson has expressed it thus:

> The central dogma [of information theory] says, "Meaning is irrelevant." Information is independent of the meaning that it expresses, and of the language used to express it.
>
> (Dyson, 2011, para. 5)

In philosophy, Bar-Hillel and Carnap (1953) called very early for a companion to communication theory that would formalise semantic information:

> The Mathematical Theory of Communication, often referred to also as Theory (of Transmission) of Information, as practised nowadays, is not interested in the content of the symbols whose information it measures. The measures, as defined, for instance, by Shannon, have nothing to do with what these symbols symbolise, but only with the frequency of their occurrence.
>
> Bar-Hillel and Carnap (1953, p. 147)

Their goal was to develop a semantic theory of information, characterised as follows:

[2] In fact, it might not take much sophistication to design an error-correcting receiver that makes use of the fact that RW rarely occurs in the messages it receives to correct COMPARW to COMPARE on purely statistical grounds. Weaver's remarks include an extensive discussion of such issues (Shannon and Weaver, 1949, §2).

> The contents of the symbols will be decisively involved in the definition of the basic concepts of this theory and an application of these concepts and of the theorems concerning them to fields involving semantics thereby warranted.
>
> Bar-Hillel and Carnap (1953, p. 148)

The resulting theory draws heavily on Carnap's earlier work defining a notion of *inductive probability*. We do not need to go through that background, though, to recognise that the formal apparatus of the theory largely consists of measures from communication theory but interpreted in terms of inductive probability rather than statistical probability. It is not clear why inductive probability is a more content-friendly concept than statistical probability. In any case, what is relevant for our purposes is that Bar-Hillel and Carnap's scepticism had an influence on philosophical understanding of the relevance of information theory for semantic content.

Dretske (1981, p. 241, n. 1) compared Bar-Hillel and Carnap's project to his own. He also picked up on Weaver's earlier remarks:

> In this sense Weaver is surely right: communication theory is not irrelevant to a specification of a signal's information content. The measurement of a one-gallon bucket cannot tell you, specifically, what is in the bucket, but it can tell you something about what *can* be in the bucket. It can tell you, for example, that it cannot contain two gallons of lemonade. And if you know, on other grounds, that the bucket contains either two gallons of lemonade or one gallon of beer, the measurement will determine, quite unerringly, the contents of the bucket.
>
> Dretske (1981, p. 41), emphasis original

This seems to be the only role Dretske sees communication theory playing in the determination of signal content. And it is an impoverished role indeed: in the bucket example, the possible contents of the bucket have nothing to do with their volume, because knowledge of

which contents are possible is antecedent. So it is with communication theory. For Dretske, all the theory can do is provide numbers. If we have knowledge of constraints with which those numbers can be matched with contents, we might learn something about contents; but the *source* of those contents is, as it were, extraneous to the theory.

Influenced by Dretske, Dennett repeated Bar-Hillel and Carnap's call for a proprietary distinction between mathematical and semantic information:

> A more or less standard way of introducing the still imperfectly understood distinction between these two concepts of information is to say that Shannon-Weaver theory measures the *capacity* of information-transmission and information-storage vehicles, but is mute about the *contents* of those channels and vehicles, which will be the topic of the still-to-be-formulated theory of semantic information.
>
> Dennett (1983, p. 344), emphasis original

Dennett (2017, §6) is still pursuing this line. The 1983 paper was on the subject of cognitive ethology, and influenced philosophers of mind, cognitive science and biology, as well as scientists working on animal communication. Section 3.5 below picks up the thread, arguing that the Shannon-semantic distinction is misleading and ought to be abandoned. The point to take away here is that Shannon's Warning escaped the context in which it was originally made. The warning applies to systems in which one set of symbols is used to represent another. The warning is that the meaning of the represented symbols are absent from the theory. It implies nothing about the meaning of the representing symbols. As we shall shortly see, the meaning of the representing symbols is indeed captured by the theory. Furthermore, the warning was not intended as a statement about other kinds of system. It was certainly not intended as a statement about biological signalling.

3.3.3 Ubiquitous Information

Another famous feature of information-theoretic mathematics is its universality. Measures such as surprisal, entropy and mutual information take statistical variables as parameters. A statistical variable is just a distribution: any distribution has an entropy, and any element of that distribution has a surprisal. Any two distributions (for which a joint distribution can be defined) have mutual information between them.

Furthermore, these measures are useful outside the context of communication theory. For example, as a way of capturing the spread of a distribution, entropy has been used as a measure of ecological diversity (Margalef, 1957; Sarkar, 2013). More pertinently, as a way of capturing correlations, mutual information has been used in domains as different as historical linguistics (Jäger, 2018) and astronomy (Pandey and Sarkar, 2017). One strange consequence is that any system can be imagined to be a communication system, simply by measuring the mutual information between two variables and interpreting the result as the transmission rate the system would have if it were being used to communicate something. This makes it seem as though communication theory itself is unable to distinguish communication systems from merely statistically correlated systems.

3.4 Overturning Irrelevance and Insufficiency

In this section I argue, first, that communication theory is not irrelevant for semantic concepts of information, and second, that communication theory quantifies the 'correctness' of signals in a way that can be reasonably considered sufficient for a philosophical account of semantic content.

3.4.1 Overturning Irrelevance

SHANNON'S WARNING leads some to argue along the following lines. The tools of information theory are irrelevant for philosophical accounts of semantic content, because those tools are agnostic to the meanings of source strings. The Irrelevance claim is usually couched in terms of a proposed distinction between Shannon information and semantic information. I address this distinction directly below, but first consider Piccinini and Scarantino's assertion of Irrelevance with respect to it:

> As we have stressed, Shannon information does not capture, nor is it intended to capture, the semantic content, or meaning, of signals. From the fact that Shannon information has been transmitted, no conclusions follow concerning what semantic information, if any, has been transmitted.
>
> Piccinini and Scarantino (2011, p. 21)

The authors intend to assert SHANNON'S WARNING. But what they say entails that *signals* in the central model need not have semantic content – and nothing they say establishes that claim.[3] The formalism of information theory is indeed blind to the meaning of the source string (if it even has one), but what the formalism captures is still *a* meaning: the content of the signal is *the identity of the original string* (and an instruction how to recover it).[4]

[3] It is difficult to interpret this passage because Piccinini and Scarantino use 'semantic content' informally while giving detailed definitions of two kinds of semantic information. However, both this passage and the rest of the article suggest that semantic information is necessary for semantic content. So if the fact that Shannon information has been transmitted entails nothing about whether semantic information has been transmitted, then it entails nothing about whether semantic content has been transmitted either. And since what is transmitted is a signal, their claim is that the fact that Shannon information has been transmitted entails nothing about the semantic content of the transmitted signal: in particular, the signal need not have semantic content at all.

[4] Here I assume a one-to-one mapping between source strings and codestrings. It is possible, however, for an encoding scheme to map multiple different source strings onto the same codestring. In that case there is still a mapping relation (or relations) between signals and source strings. It is just that the content of the signal cannot be easily expressed propositionally, in the manner of 'the content of signal x is string w'. A formal theory of mapping relations is required to capture the content of such signals. Work towards such a theory is discussed in chapter 6.

Philosophers have conflated the true claim that source strings need not have semantic content with the false claim that well-functioning signals need not have semantic content.

In section 5.5 I argue that signals in the central model bear teleosemantic mapping relations. In that section I also argue (as Millikan has argued extensively) that mapping relations are the fundamental type of semantic content. To illustrate further, consider the source coding theorem (appendix A). The theorem links entropy, a probabilistic measure, with a quantity of symbols. It clearly embodies a representational notion in some sense: ineradicably, the question that prompted the theorem and the result it offers both assume that the symbols are being used *to record the events in question*. Cover and Thomas make this clear:

> This theorem provides another justification for the definition of entropy rate – it is the expected number of bits per symbol required to describe the process.
>
> Cover and Thomas (2006, p. 115)

The extent to which "describ[ing] the process" is synonymous with the philosophers' notion of representation has not to my knowledge been asked often. I contend at the very least it is equivalent to the teleosemantic definition of semantic content.

The illicit derivation of Irrelevance from SHANNON'S WARNING misses the important fact that source strings are the *representeds* of the central model. The *representations* are the signals. The central model relies on the representational relation between them.

3.4.2 Overturning Insufficiency

UBIQUITOUS INFORMATION leads some to argue along the following lines. Informational measurements are not sufficient to specify semantic content, because they do not pick out correctness conditions. Statistical correlations, as averages, do not admit of falsity for indi-

vidual signals. All but the most permissive accounts of meaning require additional criteria that a signal must meet for it to possess content.

Birch (2014b), for example, raises this objection to Skyrms's account and suggests further criteria for correctness. Hutto and Myin (2013, p. 67) state that "covariation in and of itself neither suffices for nor otherwise constitutes or confers content, where content minimally requires the existence of truth-bearing properties"; they go on to argue that certain kinds of socio-cultural practices are necessary for the latter. As we saw in section 1.6, both Neander and Shea assert that informational measures are insufficient for correctness conditions; they each add their own criteria including functional behaviour.[5]

Indeed, it may seem to follow from my remarks in section 1.3 of the present work that UBIQUITOUS INFORMATION implies Insufficiency. There I said that the signal/cue distinction is relative to the explananda of the scientist. Informational measures such as mutual information do not distinguish signals from cues. Such measures can therefore only be used as part of an explanans whose explanandum does not distinguish signals from cues. However, a theory of semantic content *should* distinguish signals from cues. A theory of content ought to tell us why signals possess correctness conditions and cues do not. Another way to say this is informational measures are agnostic to the functions of the items they measure. Since functions (I claim) are the heart of semantic content, those measures cannot capture semantic content.

We can reject the argument as follows. Communication theory does distinguish between

[5]It is less clear how Millikan takes her version of teleosemantics to relate to information theory. Millikan (2017, p. 111) notes some differences between her definition of "natural information" and what she takes the information-theoretic concept to be, but her aim seems to be distinguishing her account from that of Dretske, rather than denigrating information theory.

mere statistical correlations and those that can be used for signalling. A statistical correlation would not support an encoding scheme. Changing the value of X would not lead to a change in Y. The central model defines a directional causal link between the two. The existence of a *causal* relationship between X and Y is required for Shannon's theorems to apply. In short, there is a difference between *being able to learn about X from Y* and *being able to use X to communicate something to Y*.

Furthermore, signals in the central model bear teleosemantic mapping relations. They are correct to they extent that they accord with the encoding procedure with which the system was designed. For example, consider a system that transmits outcomes of coin tosses in accordance with the code $H \rightarrow 1,\ T \rightarrow 0$. If a 1 is transmitted when the coin comes up tails, the signal is false.

As a statistical average, mutual information – transmission rate – measures the extent to which signals are inaccurate. In a system in which no signals map, transmission rate is zero. In a system in which every signal maps perfectly – there is no noise, and the encoder makes no mistakes – transmission rate takes its maximum value $H(X)$. Mutual information is a measure of the extent to which signals bear mapping relations, or alternatively the accuracy of the relations they bear. This is so despite the fact that mutual information can also be used to measure statistical correlations outside communication systems. The correlation between a cue and its signified can be measured by mutual information, but then it is not being used as a measure of accuracy.

What is more, communication theory provides tools for quantifying the severity of misrepresentation (Martínez, 2020). As explored in chapter 6, rate-distortion theory describes trade-offs between signalling effort and accuracy of receiver behaviour. Such results would

be impossible if there were not some sense in which signals in the central model can be false. Sceptics might wish to argue that this is not the *same* sense of falsity required for a theory of content, but since rate-distortion theory has been all but ignored in the relevant philosophical literature, they would have to present novel arguments in order to do so.[6]

Proper function is not made explicit in communication theory. Nonetheless, as an engineering discipline the systems it deals with are implicitly functional. Teleosemantics applies, and central model signals have truth-conditional contents. The extent to which a given signal is false can be quantified by considering the extent to which receiver behaviour deviates from what the signalling system was designed to support.

The next section uses these considerations to argue that a popular distinction between Shannon information and semantic information is misleading, and should be replaced by a distinction between cues and signals.

3.5 The Shannon/semantic distinction

In this section I argue that we should reject a popular distinction between Shannon information and semantic information. The explanatory work currently assigned to the Shannon/semantic distinction is best performed by the more clearly understood distinction between signals and cues. As described in chapter 1, signals are transmitted and received by codesigned entities, while cues are received by an entity not codesigned with the sender. The conceptual differences between these two categories, as well as the range of formal tools required to analyse them, are well understood and well supported.

[6]Of course, plenty of philosophers argue that teleosemantic mapping relations do not provide falsity in the manner required for a theory of content. Presumably, Insufficiency is tenable once teleosemantics is denied.

Recall, cues are signs whose sender is not codesigned with its receiver. A receiver benefits from learning how to respond to the stimulus, but the source of the stimulus does not benefit – either because it has incompatible interests with the receiver or because is not an agent at all. In contrast, signals definitionally require a cooperative relationship between sender and receiver. Both benefit from coordinating their behaviour with the use of a sign, produced by one and perceived by the other. I shall argue that the cue/signal distinction is useful and accords with both theory and practice in natural science. Meanwhile, the Shannon/semantic distinction is inspired by a confusion about communication theory, and obscures fruitful relationships between models of communication and strategic behaviour.

The Shannon/semantic distinction (Piccinini and Scarantino, 2011, §§4.1-2) (Godfrey-Smith and Sterelny, 2016, §§2-3) has at least two points of origin. It is firstly a mutated form of an earlier distinction between natural and intentional meaning, which may be traced back at least to Brentano and found its clearest statement in Grice (1957). Prompted by Dretske (1981, 1988) the distinction took centre stage in the teleosemantic literature of the 90's (Millikan, 2001). The original distinction is still hard at work in Millikan's teleosemantics (Millikan, 2017, §§11-12), but its mutated form is misleading. A second source is Bar-Hillel and Carnap's clarification of "information" as it appears in communication theory. They distinguished the mathematical quantity from the semantic notion which is of interest to philosophers (Bar-Hillel and Carnap, 1953). Dretske compared Grice's approach, as well as that of Bar-Hillel and Carnap, to his own project (Dretske, 1981, pp. 241-2, n.1 and n.10). Soon after, the "still imperfectly understood" distinction was cited by Dennett (1983, p. 344 col. 2) and picked up by Krebs and Dawkins (1984, §§4.1-2), whence it found its way into the behavioural ecology literature and prompted ongoing scepticism about the use of infor-

mation theory in the study of animal signalling (Owren et al., 2010; Pfeifer, 2006; Sarkar, 2013).

The Shannon/semantic distinction follows from the Irrelevance claim. It entails that whatever semantic information is, it must be something richer than Shannon information, something that requires a different formal framework. This distinction has become so widely accepted that the Stanford Encyclopedia entry "Biological Information" is currently organised around it (Godfrey-Smith and Sterelny, 2016). I have already argued against the irrelevance claim in section 3.4. In this section I offer further arguments against the coherence of the Shannon/semantic distinction.

I am going to argue by analogy with the property of *volume* and relational property *weight*. Because those properties are well understood, I hope to show that (often implicit) justifications for the distinction are bogus. I will draw on an argumentative pattern that can be constructed for information, volume and weight. In the case of information, the argument is as follows:

> The **information** of a signal cannot tell you what signal it is. Many different signals can be measured using the same unit, bits. Since different signals can have the same number of bits, measures of information cannot tell you the referent of a signal. Further, the capacity of a channel – the maximum information that can be transmitted through it – cannot tell you what signals will be passed through it, or which states of the world will be the referents of those signals. Therefore, there are two concepts of information: SHANNON INFORMATION, which concerns quantities, and SEMANTIC INFORMATION, which concerns referents.

An argument along these lines seems to be what Scarantino has in mind in the following passage:

> [W]e cannot fully capture the informational content of a signal by merely listing the amounts of incremental information being carried. Informative signals do not tell us just *how much* probabilities have changed; they also tell us what are the *states of affairs* that had their probabilities changed. On Skyrms's account, two signals that carry the *same amount* of incremental information with respect to four *completely different* states of affairs would have the *same informational content*, captured by, say, the vector $< 1,\ 3.4,\ -\infty,\ 1 >$. This is unacceptable.
>
> Scarantino (2015, p.429, emphasis original)[7]

To rebut the argument, it will be worth exploring a couple of fanciful analogies.

Consider a mathematical theory of spatial volume. This theory provides units of measure for volumes of objects. These units are commensurate with cubic metres, but in order to make the analogy vivid we will call them *Tufnels* after the fictional scientist who originated the theory. All physical objects can be associated with a measure of Tufnels. Your car might be 1400 Tufnels in volume while your house is 35 kilotufnels. Objects can in principle be compared in this way – you can say that your house is greater in volume than your car – but it's not obvious why such a comparison would be useful (unless you were trying to fit your car inside your house). An extension to the theory allows engineers to measure the amount of liquid that can be passed through an intermediary like a pipe. Importantly, the measure of *Tufnels per pipe segment* does not specify what is being passed through the pipe. But we can compare different pipes in order to see which will deliver liquid faster between two points. We can measure the capacity of different pipes, but these measurements cannot tell us what is

[7] Scarantino is right to say that statistical content does not define correctness conditions but, as the following analogies are intended to show, he is wrong to say that statistical content does not tell us what are the states of affairs whose probabilities changed. Scarantino seems to assume that if a sign is *about* its signified, it must have that signified as a correctness condition. Skyrms (2010) and Isaac (2018) also appear to share this assumption; it is precisely what I deny.

being pumped through them. An engineer may use Tufnel's mathematics to choose the right pipes for my house, but she cannot guarantee that the water board will not pump fluoride through them. We tentatively conclude that there are two concepts of volume. First, there is Tufnel volume, a concept associated with mathematical measures which allows engineers to do useful things like recommend pipe installations. Second there is what we may call spatial volume, a concept that *identifies* objects with spatial extension. Tufnel's theory cannot say what the objects of its measurements are: it can only measure them. We therefore need a supplementary theory of volume, one that delivers the identity (whatever that means) of spatially extended objects. In short:

> The **volume** of an object cannot tell you what object it is. Many different objects can be measured using the same unit, tufnels. Mathematical theory can tell us about quantities of volume, but it cannot tell us about what objects those quantities are volumes of. Further, the capacity of a pipe – the maximum volume that can be put through it – cannot tell you what objects will be placed inside it. Therefore, there are two concepts of volume: TUFNEL VOLUME and SPATIAL VOLUME.

The argument is fallacious. Nobody expects mathematical measures to *identify* the objects measured. If you measure something, you already know what you are measuring. Further, the measure of capacity is an extension to, or useful application of, the original measure of volume. Not only does a volume measure have nothing to say about the identity of the object with that capacity, it is silent about the objects that can be put inside that thing – *except* with regard to the property it is designed to tell us about, namely volume. The fact that it is silent about the identities of objects is a feature, not a drawback, of the volume measure. The concept SPATIAL VOLUME therefore has nothing to do with volume per se. It is a chimerical

concept.

It is not clear what the counterpart theory of volume would be. A theory of the identity or categorisations of objects? But what does this have to do with their spatial extension, if it has nothing to do with the magnitude of this extension? Part of the confusion here stems from the following legitimate observation. In the case of pipes, there can be a measure of capacity without specification of what is flowing through the pipes (compare SHANNON'S WARNING). This capacity, of course, is a rate: it says how much of something as yet unspecified can flow through the pipe in a given time period. But note two crucial points that are missed by this description. First, the measure in question is *capacity*, not volume. It tells of how much something can contain, not how big something is. It is a derivative measure. Second, there *is* an object associated with this capacity: the pipe itself. We have abstracted away from what might be flowing through the pipe, but we are not thereby providing a measure ignorant of physical objects: the pipe itself is an object, and the measure is associated with it. The point comes through more forcefully when we talk about buckets rather than pipes. A bucket has a capacity, which states how much of something can fit into it. We can specify a bucket's capacity without saying anything about what is or could be placed into the bucket. This is no way entails that there are two concepts of volume, one that is agnostic to objects and one that identifies them.

Incidentally, measures of capacity are not *entirely* agnostic as to what can go inside the bucket. A bucket with a capacity of 1000 cubic centimetres cannot hold a steel rod of cross-sectional area 10cm^2 and length 100cm. It depends on the shape of the bucket, which might be shallow and squat. The rod, which is rigid, may poke out over the top of the bucket. In general, what is assumed by measures of capacity in both buckets and pipes is that the object

they are to hold is liquid or gas. It is also typically assumed that the object carried is not corrosive to the material out of which the bucket is made. A litre of hydrofluoric acid has the same volume as a litre of water, but only one of them can be reliably contained in a metal bucket with a volume of one litre.

The analogy with volume is silly in part because volume is a one-place property. A better comparison would consider relational properties. Let us explore one such.

Consider a mathematical theory of weight. This theory provides units of measure for weights of objects, called *Newtons* after the scientist who originated the theory. All physical objects can be associated with a measure of Newtons. Your car might be 1400 Newtons in weight while your house is 35 kilonewtons. Objects do not have weight intrinsically: it is a relational quantity that depends on the planetary body on which the object is found. An extension to the theory allows engineers to measure the amount of weight that can be lifted by a platform like an elevator. Importantly, the measure of *Newtons per elevator* doesn't specify what is being lifted. Nor, initially, does it say how fast the object moves: we would need to add a measure of how many metres per second the platform could traverse, in order to obtain a measure of Newtons per second. Once we have this measure, however, we can compare different elevators in order to see which will deliver objects faster between two points. We can measure the capacity of different elevators, but these measurements cannot tell us what is being lifted by them. An engineer may use Newton's mathematics to choose the right elevator for my office, but she cannot guarantee that burglars will not use them. We tentatively conclude that there are two concepts of weight. First, there is Newton weight, a concept associated with mathematical measures which allows engineers to do useful things like recommend elevator installations. Second there is what we may call spatial weight, a

concept that *identifies* objects with mass. Newton's theory cannot say what the objects of its measurements are: it can only measure them. Furthermore, the theory cannot tell us which planetary body the object has weight with respect to. We therefore need a supplementary theory of weight, one that delivers the identity (whatever that means) of objects with mass. In short:

> The **weight** of an object cannot tell you what object it is. Many different objects can be measured using the same unit, Newtons. Mathematical theory can tell us about quantities of weight, but it cannot tell us about what objects those quantities are weights of, nor with respect to which gravitational body they are measured. Further, the capacity of an elevator – the maximum weight that can be put into it – cannot tell you what objects will be placed inside it. Even further, weight measurements cannot tell you which gravitational body the weight is being measured with respect to, nor, in the case of the elevator, which planet the elevator will be installed on. Therefore, there are two concepts of weight: NEWTON WEIGHT and SPATIAL WEIGHT.

What was said about the concept of SPATIAL VOLUME applies equally here. Mathematical quantities are not supposed to identify objects. The weight of an object tells you neither what object it is nor what planetary body its weight has been measured with respect to. The weight-capacity of a lifting platform does not tell you which platform it is, nor what objects it will be used to lift. Again, this is a feature rather than a drawback. Again, the concept SPATIAL WEIGHT is chimerical.

What is different about weight is its relational nature. It is a more suitable analogy for the case of signals. With signals, what we want to know is the identity of the state of affairs that is the signal's referent. This is akin to wanting to know the identity of the planet that an object is weighed with respect to. But if you weigh an object, you already know what planet

the measurement is being made with respect to.

Again, the sceptic might complain that communication channels are agnostic to the meanings of signals that flow through them. But this is to say nothing more significant than that elevators are agnostic to the objects they lift. The sceptic might retort that communication channels, no matter the size, can transmit meanings that are extraordinarily significant to their receivers. For example, a 1-bit channel may carry a signal interpreted as 'attack' or 'retreat'. A single signal sent through this channel may launch a thousand ships; surely this is a magnitude of significance not captured by communication theory? But this is to say nothing more than that an elevator on the moon can carry much more than an elevator on earth. The significance of a signal is relative to its users, just as the weight of an object is relative to its planetary body.[8]

3.5.1 The origin story of Shannon information

It is possible to reconstruct my opponent's argument more charitably.[9] In understanding physical properties, there are (at least) two kinds of approach we can pursue. One is the *measurement approach*: we develop mathematical measures and engineered tools to allow us to measure the property in question. Another is the *analysis approach*: we attempt to determine how the property is manifested, by investigating the more fundamental objects and properties that give rise to it. In the case of weight, the measurement approach was most famously spearheaded by Newton, but the (correct) analysis had to wait until general relativity and accounts of fundamental particles like the Higgs boson. In the case of semantic content,

[8] I have mixed up the analogy a little here. Initially, I treated the gravitational relation as analogous to the sign-signified relation. Here I treat it as analogous to the sign-user relation. I do not think this really matters; the point to bear in mind is that measures of relations already have their relata specified.

[9] An anonymous examiner suggested the following account.

the measurement approach we are considering in this chapter is the theory of information founded by Shannon. The analysis approach for semantic content belongs to the discipline of analytic philosophy.

The fact that there are these two kinds of approach does not by itself entail that there are two kinds of properties. But once a certain measurement approach reaches maturity, it could turn out that the property it tracks is *different* from that which is the target of the analysis. That is the origin story of Shannon information: it is measuring something real, but what it measures is not what was previously informally referred to by the term 'information' (and what philosophers now call semantic content).[10] The same might have been true of Newton's approach: it might have turned out that what he ended up measuring was different from the property we now call gravity, and which is analysed by general relativity and particle physics.

My opponent's arguments, as I reconstruct them in this section, move from claims about features of the measure to the claim that there must be two distinct properties involved. It is certainly possible that the measurement approach to a certain property could, over time, diverge from the original property it sought to measure. Then there would be two distinct properties in question. One would be tracked by the original analysis approach, the other would be a novel property accidentally discovered as a result of refinement of the measure. The difficulty with claiming that this has happened is in establishing that what is being measured is different from the original target property. My opponent claims to have established this for Shannon information. The analogies above supposed to show that that claim cannot possibly be established in the way my opponent intends.

[10] Shannon probably never considered himself to be trying to measure semantic content. The point of the origin story is just to show how my opponent seeks to justify the claims of irrelevance and insufficiency.

It is not possible to move from the claim that 'different signals can have the same number of bits' to the claim that 'there are two different concepts of information'. It is not possible for the same reason that it is not possible to move from the claim that 'different objects can have the same weight' to the claim that 'there are two different concepts of weight'. My point is not (just) that communication theory provides tools for measuring the relations that teleosemantics claims ground semantic content. My point is that my opponent cannot possibly have established the counter-claim. If they could, they would be able to establish similar claims for volume and weight.

Scarantino, for example, objected that statistical content vectors omit the referents of signals because they contain only numbers. Two signals bearing quantitatively similar probabilistic relations to different states of affairs could have the same content. In much the same way, two objects bearing quantitatively similar spatial and mass-ratio relations to different gravitational bodies could have the same weight. Pursuing the analogy, we might imagine 17th century metaphysicians interested in the relation between everyday objects and the earth, or between the earth and the sun. Suppose they learned of Newton's work, but objected that although he managed to quantify gravitational force, something about the relation of interest was omitted from his calculations, because two distinct objects might have the same weight. Such a view would be at best unfair, at worst absurd.

Scarantino believes statistical content vectors ought to contain an explicit term that differentiates one event space from another. But that is not how the relevant mathematical terminology works. Each content vector has an already defined event space W, just as each weight measurement is made with respect to an already specified planetary body. Specification of a sign's statistical content vector ineliminably includes its referent. I suspect that if

Scarantino's concerns were valid, it would *never* be possible for a specification of content to pick out the state of affairs that is the sign's referent. After all, we must construct a definition of content using only the terms available in our language. Following Scarantino, we might include a phrase such as 'eagle present' in our content specification. But such terms do not *by themselves* indicate the state of affairs; we require, by Scarantino's logic, an extra piece of terminology that gives the correct interpretation of the term. But we would then require an interpretation of the interpretation, and so on.[11] Scarantino's argument is predicated on a confusion about what must be explicitly versus implicitly stated in a formal definition.

Bringing us back to the target of the analogy, the central model embodies many assumptions corresponding to those of fluid engineers installing water pipes or builders installing elevators. Not everything may be transmitted, because there may not be a suitably efficient encoding. There is even a correspondence with corrosive material. Depending on the range of behaviour of the receiver, there are strings that cannot be passed in full. For example, consider a string with an instruction in the receiver's language: "Destroy this channel. Print '456'." Assuming a memoryless receiver, the second part of the instruction will not be carried out, because the first part is 'corrosive'. It may superficially have the same information content as the message "Print '123'. Print '456'." But its effects are different.

I remain unconvinced that the distinction between Shannon and semantic information is tenable or even coherent. I will now demonstrate that the job vacated by that redundant distinction is best performed by the cue/signal distinction.

[11] This regress seems related to the Kripke-Wittgenstein paradox (Wittgenstein, 1953, §198ff) (Kripke, 1982).

3.5.2 Cues and signals

As we saw in chapter 1, cues and signals lie on a continuum. Just as there can be degrees of adaptation leading from disposition to behaviour, there are degrees of coadaptation from joint disposition to joint behaviour. A sign that mediates an interaction becomes more signal-like, on this definition, when production of and response to the sign become coadapted behaviours. The sign is more cue-like as production and response are less coadapted.

A cue has no function *qua* cue. Whatever 'sends' a cue, if it has a function, is not the same function as that for which the cue is used. So for example, a sign sent between two cooperating entities is a signal for them, but is a cue for an eavesdropper. The contemporary view seems to have confused cues for Shannon information, which prompts the belief that informational quantification can only measure cues. Given that anything could in principle become a cue if there was an agent who could make use of it, Shannon information is thought to be an incredibly broad resource, and as a result virtually useless for biological theorising (Godfrey-Smith and Sterelny, 2016; Kolchinsky and Wolpert, 2018; Shea et al., 2017).

Contemporary scholars, I suggest, often unknowingly aim for the cue/signal distinction. The popular claim is that Shannon information captures statistical correlations, and this alone cannot distinguish cues and signals (Godfrey-Smith and Sterelny, 2016, §2) (Owren et al., 2010, pp. 772–3). However, the term 'Shannon information' promotes confusion here. It is true that the information measure employed by statistical decision theory is typically applied to cues and not signals. It is also true that this measure is derived from Shannon's work (Shannon, 1948b) which is itself a continuation of Hartley (1928). But communication theory has other formal tools beyond measures of correlation. It has other commitments besides the condition that signals covary with their sources. Communication theory is not simply

statistical decision theory, though it is closely related. The mistake is likely promulgated by the association of Shannon's name with the quantity used in decision theory. It would be best to retire the misleading label, and stick to distinguishing the explanatory roles of signals and cues. One of them bears mapping relations while the other does not.

The sceptic might argue that Shannon information is just another name for statistical content. As I myself claim, statistical content does not capture correctness conditions. We need another formal tool, and I endorse functional content (chapter 6). One might wonder whether this is the distinction I have been claiming is incoherent. It would admittedly be a partial victory to cement 'Shannon information' as statistical content and 'semantic information' as functional content. But I am reluctant to endorse this move, for two reasons. First, association of Shannon's name with a mathematical term that is known to be too weak to capture correctness conditions leads many to assume communication theory has no more tools to offer. I have already cited Neander, Shea, Dennett, and Piccinini & Scarantino as holding this view or something dangerously close to it. If they all can be misled, something radical needs to change, and I suggest removing Shannon's name from the proceedings. Second, neither statistical content nor functional content are explicitly defined in communication theory, although both are implicit. We would be equally justified in calling functional content 'Shannon information', glossing it as 'what a signal is supposed to indicate about the source'. In fact, there is probably *better* justification for associating Shannon's name with functional content than statistical content: signals in the central model have functional content, while other statistical relationships investigated with information-theoretic tools (with which Shannon was less concerned, if at all) do not.

3.6 Objections

3.6.1 Scaling down: mapping relations are not semantic content

Objection. A recent trend seeks to distinguish two concepts I treat as equivalent. The distinction, advocated by several authors including Price (2008, §5), Hutto and Myin (2013, p. 67), Rescorla (2013) (who cites Burge (2010) as inspiration) and Lean (2014), runs as follows. Simple signalling systems carry information in the guise of *reliable correlation* ("functional isomorphism", "Shannon information") – tokens that correspond to worldly states in a manner sufficient for successful behaviour. But correlational information is to be distinguished from the much richer notion of *content*, which is characterised by truth conditions.

Response. There is far more to say about this distinction and its motivations than can be addressed here. I have already remarked on the unfortunate term "Shannon information" (section 3.5). The distinction lies at the heart of a family of objections to teleosemantics. Roughly, these objections state that the theory is unable to account for the content of many familiar mental states like beliefs and desires. But these objections rely on an unspecific concept of semantic content. In particular, scholars who press the objection typically assume a distinction between the content of mental states and the content of biological signals, without sufficiently characterising that distinction (Millikan, 2013b). In particular, the 'rich' notion of content that supposedly belongs to mental states is not defined in enough detail to distinguish it from the 'poor' notion that teleosemantics ascribes to biological signals. As a result, these proposed objections to teleosemantics are as yet inconclusive.

Nonetheless, these are pressing concerns for those who wish to employ teleosemantics in philosophy of mind and cognitive science. Fortunately, since we are here only dealing

with biological explanation, our task is much less complex. We need only make plausible the claim that the distinction between Shannon information and semantic information is chimerical in communication-theoretic applications. Consider, for example, a definition of semantic information in terms of truth conditions. As custom has it, semantic information is *about* something – can be true or false of the objects or events that are its topic. On the contrary, Shannon information need not be about something. This characterisation ought to be enough to distinguish the two concepts, so why do I claim the distinction is unreal?

It cannot be the case that semantic information is about something whereas Shannon information is not. The premise of the current objection is that simple signalling systems carry *only* Shannon information. This is equivalent to saying they exploit reliable correlations in order to achieve success. But if a system is exploiting a reliable correlation, it is dealing with some accessible token that bears a relation to some other inaccessible process or object. This relation *is* 'aboutness', on the teleosemantic account.

The objection, then, is not so much that biological signals do not have 'aboutness', but that their 'aboutness' is not sophisticated enough to be *content*. Without sufficient characterisation of this latter term, however, the dispute risks being merely verbal. Different scholars use the term 'content' to mark different distinctions. In this project, I outline what I take to be important distinctions in the context of biological explanation. The issue of labelling the different entities should be tackled *once they have been firmly distinguished*. For example, one might say that a token is a signal only when it is produced for the purposes of being used. This would not prevent others from using the term 'signal' more broadly. For example, preserved features that can be used to infer historical processes are sometimes called 'signal' in phylogenetics, though in my terminology they would only be cues (whose receiver is a sci-

entist). Similarly, one might say that a signal is a representation only when it is sufficiently decoupled from use. This would not prevent others from using the term 'representation' more broadly – in particular, using it as a synonym for signal, as Millikan does.

If all this is agreed, the objection states that there is some firm, explanatorily relevant distinction between the aboutness possessed by biological signals and the aboutness possessed by states that are the traditional subject of philosophical theorising – beliefs, desires, and other mental states. So be it: I need not respond to that objection here (though I offer relevant comments in section 5.4). This is a work of philosophy of biology, not philosophy of mind. I am not here concerned with defending teleosemantics against claims that it cannot serve its original purpose, to account for the semantic content of thoughts and human natural language. I am only concerned to show that it adequately interprets applications of communication theory in biological contexts. I am satisfied that by drawing important distinctions, we can see the proper application of the theory to this task. Using the term 'semantic content' to refer to teleosemantic mapping relations may cause problems in a wider theoretical setting. Within the scope of this project, however, I have been clear about how the term ought to be interpreted.

3.6.2 Monoculture: Shannon information is semantic

Objection. Brian Skyrms (2010) tells us that information is everywhere, and has semantic content regardless of whether it is used. In particular, semantic content can be defined in terms of objective probabilities. Coevolved senders and receivers use signals to coordinate behaviour, but they are doing nothing more complicated than producing a resource that is already prevalent in the world. Alistair Isaac (2018) gives this claim real force, by expanding

Skyrms's formalism to give a rigorous semantics for Shannon information. According to Isaac, a semantics must accord with the following criteria:

> I take the constitutive feature of a semantics to be that it assigns a unique, evaluable formal object to each element in a set that characterizes all and only the content conveyed by that element – intuitively, what it 'says about the world'.
>
> Isaac (2018, p. 6)

A semantics for a system is a method of determining the meaning of each element in the system. Isaac claims statistical content counts as a semantics for Shannon information.

Scarantino (2015) pursues a similar line. He questions the explanatory distinction between signals and cues, in effect defining falsity in terms of receiver expectations. On this view, both signals and cues can be false, and their falsity explains failure in the same way. Like Isaac, Scarantino describes a formalism that can be interpreted as a semantics for signals and cues. Unlike Isaac (and unlike Skyrms), he is non-committal on the interpretation of probabilities underlying the formalism. While Isaac follows Skyrms in being an objectivist about probabilities, Scarantino's approach is ecumenical: his formalism holds across different interpretations of probability.

In all, then, a single notion of information is sufficient to do explanatory work in biology. In contrast to my distinction between signals and cues, these authors claim that the 'aboutness' borne by both is equivalent. Skyrms, Isaac and Scarantino claim that cues can be false, on the grounds that their semantics assigns content to them.

Response. When modelling a sign-using system, whether or not signals differ from cues, and how they do if they do, depends on our explanatory goals. Signals and cues are theoretical entities, designed to participate in explanations of success, failure, and surface behaviour.

But the objection seems to draw an unreasonable conclusion from a reasonable premise. In section 1.3, we saw that the distinction between signals and cues collapses when we ignore the agency of the sender. Similarly, communication-theoretic problems collapse into decision-theoretic problems when the design of the sender is ignored. The objection seems to require that such a perspective is the *only* one – or at least a privileged one – from which to explain behaviour. And such a position, I maintain, is untenable. Explaining signalling behaviour sometimes requires treating the entire system as a functional unit. Explananda in such cases may include sender and receiver behaviour or morphology. It may also include the form of the signal itself, in cases where the signal is physically distinct from the sender. Such explanations do not exist for cues. The form of a cue is never explained by its use.[12]

Furthermore, I agree that even cues have a 'semantics' in Isaac's sense. For him, a semantics is an association of a formal property with a set of signs. Establishing such a formalism is not sufficient to establish that cues can be false, and therefore that the distinction between cues and signals is explanatorily irrelevant. Isaac seems to think that having a semantics in his sense *entails* the possibility of falsity. But simply assigning a semantics to a system does not entail that its signs can be false. The fundamental problem of naturalistic intentionality is determining how signs can have content whose satisfaction conditions do not obtain. Assigning formal objects to each sign in a system, and calling those formal objects 'content', does not in itself ensure that those contents can be false.

[12]In addition to signals and cues, there are other information-bearing items that do not obviously fall into either category. Camouflage is a morphological trait designed to *prevent* receivers achieving their goals. As such, its physical form is in part explained by considerations of biological function. In this way, camouflage differs from cues. But camouflage also differs from signals in that it cannot be explained by reference to coadaptation (at least not when 'coadaptation' implies a common evolutionary goal). So my comments in the main text should not be read as an exhaustive account of the explanatory distinctions relevant for evolutionary interactions.

Rejoinder. Let us accept there is an explanatorily relevant distinction between signals and cues. Nevertheless, it may be objected that Shannon information is sufficient to do the explanatory work biologists require of signals. Lean (2014) argues that success and failure of signal-driven behaviour can be attributed to the informational properties of a signal without appeal to mapping relations. For Lean, the distinction I drew in chapter 1, between mapping relations borne by signals and correspondence relations borne by cues, is bogus. There is one kind of relation, and what distinguishes signals is that they can sometimes fail to bear it. Signals can fail to carry the information it is their function to carry (and this distinguishes them from cues), but the kind of information they bear is the same as that of cues. This accounts for the special explanatory role of biological signals.

Response to rejoinder. Teleosemantics provides a definition of falsity that rests on the universally agreed point that signals have functions, whereas cues do not. Lean believes this distinction does not entail a distinction in the *kind* of information a sign bears. He assigns explanatory burden to cues and signals by asserting that they carry the same kind of information, but that signals can sometimes fail to do so. However, this conclusion relies on a crucial vagueness in the term 'information'. In Lean's parlance, a signal failing to carry information means a *token* signal failing to bear the requisite correspondence relation. But the term 'carrying information' typically refers to bearing a correlation, which is a statement at the level of *types*. Our distinction between correlations, correspondence relations and mapping relations carefully distinguishes terminology across two levels. But Lean's term 'carrying information' conflates them. As a result, his assertion that signals have the function of conveying information inherits the same vagueness.[13] For a token sign to have the

[13] Strictly, proper functions are causal effects. A signal cannot have as its proper function to bear a relation

function of corresponding, there must be a *token* correspondence relation it is supposed to bear to its signified. Lean requires that this token correspondence is the same kind of thing for signals as for cues, in effect denying the grounds of the distinction laid out in chapter 1. Again, this ignores the fact that signals bear the relations they do in virtue of being produced in the proper way. If we ignore the sender's function, signals simply become cues in our model of the situation. Then they either bear the correspondence relation or do not: there is no room for falsity. The only way it can be possible for a signal to be subject to a standard which it can fail to achieve is if there are rules governing Normal performance of its proper function (see chapter 4). Those rules specify which relation(s) the signal must bear in order to perform their proper functions Normally. These relations are mappings because they derive from functional specification of senders and receivers. In contrast, the relations that characterise cues are correspondences because they are derived from a completely different source: the causal structure of the world irrespective of their receiver. Ignoring this distinction is only valid when we are ignoring the role of the sender's design in explanations of receiver behaviour or success. And while biologists do this some of the time, there are times when they do not.

(unless it is a relation to a causally downstream event), though it can be required to bear a relation in order to perform its function in a Normal way (see chapter 4 and Ryder et al. (2013, pp. 37–40)). Lean must therefore work with a different notion of function than I assume throughout the thesis. I do not think this affects the point I make here.

3.6.3 Disanalogy: The analogy between Shannon information and other mathematical quantities is fallacious

Objection. Shannon information is not like weight. For one thing, the weight capacity of an elevator is different from the information capacity of a channel. If you put too much weight in an elevator, it will fail to move or the cable will snap. No analogous thing will occur with communications channels. Indeed, it is not even sensible to imagine putting too much information into a channel – such a thing is incoherent.

Response. The objection fails because there is a perfectly coherent notion of transmitting too much information through a channel. Indeed, channel coding is in part a process of regulating the amount of information transmitted. Good channel coding throttles transmission, ensuring not too much is sent per unit time. The key to understanding the situation is to remember that capacity is a *rate*. It tells how much information *per channel usage* or *per unit time* can be transmitted. When one attempts to transmit at a greater rate, one inevitably fails to reconstruct the message perfectly.

There are also more familiar issues with transmitting (or in this case receiving) too much information per unit time. Denial of service attacks occur when an internet service is flooded with bogus requests, overloading the system and preventing it from dealing with legitimate operations. Computer systems are rate-limited just as communication channels are, and malicious actors can jam systems by effectively forcing them to attempt to deal with too much information at once.

We should understand the problem of attempting to transmit too much information in terms of functional performance. Transmission rate measures how much functional improve-

ment one can enjoy by conditioning on signals sent through the channel. Suppose a sender and receiver acted as though a signal were more accurate than it really was. They might fail to strategise with respect to noise in the channel. Their behaviour could be construed as an attempt to transmit too much information through the channel, resulting in imprecise behaviour. Their behaviour would fail to the extent that attempted transmission exceeds capacity.

Rejoinder. There is a difference between transmitting too much and attempting to transmit too much. In the elevator case we really do put too much weight in the elevator, and it collapses. In the channel case we don't even get to transmit the information, by definition.

Response. We ought to read 'attempting to transmit too much information' as analogous with 'attempting to lift too much weight'. You can put too much weight in an elevator but it won't then function properly. What is being attempted is a certain level of performance. In both cases, capacity specifications tell you what level of performance you can hope to achieve. Attempting to exceed this will result in failure.

3.6.4 Priority: semantic information came first. Whatever Shannon is measuring, it isn't information

Objection. In everyday parlance, the 'information' in a signal refers not to a quantity but to the identity of the state of affairs that bears a certain relation to it. If anything, it was poor judgement on Shannon's part to adopt that term as the name of a mathematical measure. The original concept INFORMATION really *is* about the identity of objects (or at least states of affairs). This concept cannot be chimerical (or at least not for the same reasons SPATIAL VOLUME and SPATIAL WEIGHT are) because it pre-dates the mathematical definition. The

mathematical concept (if mathematical measures can be said to be concepts at all) is simply confusingly misnamed, and would better be called something technical like 'surprisal' or 'statistical content'.

Response. Such a view is not totally at odds with my argument in this chapter. Sometimes it is appropriate to compare mathematical measures with concepts, such as volume, weight, and temperature, especially in discussions of the conceptual advances that mathematical precision makes possible. What seems to have gone wrong in this case is that the mathematical definition cannot sharpen the original concept INFORMATION. That concept was already too vague to be analysed by a single measure. What was required, first, was a teleosemantic theory to sharpen the concept of information by replacing it with mapping relations and correspondence relations, followed by a mathematisation of teleosemantics to enable quantification of those relations. Perhaps the irrelevance claim is a consequence of the accidental fact that communication theory emerged before teleosemantics.

Whatever the reason, I maintain that distinguishing the state of affairs that a signal is about from the quantity of information in the signal is illicit. The irrelevance claim states that quantities have nothing to do with referents. Such a claim cannot be true. Quantities of information involve referents in the same way that quantities of weight involve planetary bodies with respect to which weight is measured.

3.7 Conclusion

What can communication theory say about content? A great deal, if we take sender-receiver teleosemantics seriously. Informational measurements in communication systems are given

significance by the proper functions of signallers. Those measures quantify the accuracy of mapping relations.

The relevance of communication theory for the philosophical problem of semantic content has been obscured. The irrelevance claim is widespread, superficially plausible, and entails a distinction between Shannon information and semantic information. Once we ground information in function, we can better appreciate the relationship between communication theory and teleosemantics.

Chapter 4

Mapping relations

4.1 Introduction

How can appealing to the *content* of signals help explain behaviour? Why should this kind of explanation be preferred to a purely mechanical description of a signalling system?

The usual approach to these questions, and the approach I shall endorse, is to show how appealing to content can offer greater explanatory purchase than mechanical descriptions. Explanatory purchase is a fundamental aspect of philosophical approaches to representation (Shea, 2018, p. 205ff), (Ramsey, 2007, p. 27). The problem for the naturalist is to define a naturalistically respectable category that scientists can appeal to in explaining behaviour, and which enables an improvement over mechanical explanation. Since content is metaphysically suspicious, an increase in explanatory purchase over purely mechanistic explanation would help to provide justification for attributing content to physical vehicles. The greater the explanatory purchase, the greater the motivation for accepting content as part of the natural order.

Understanding how content offers explanatory purchase requires understanding its explanatory role. According to teleosemantics, the relevant explanatory role should be understood in terms of a certain kind of relation that may be borne between a signal and its truth condition. When this relation holds, the signal can aid the success of its receiver in performing its proper function. When the relation fails to hold, the signal cannot aid success in this

way. Typically, when the relation does not hold, the receiver will fail; any success it enjoys as a consequence of conditioning its behaviour on the signal will be accidental. Relations that play this kind of explanatory role are called **mapping relations**.

This chapter will spell out in detail the pattern of explanation just outlined. Mapping relations are characterised by the role they play in explanations of successful performance of a proper function, and my aim in this chapter is to treat **interventionism** as an explanatory framework for this kind of explanation. Since the explanations provided by interventionism are couched in terms of causal models, it is necessary to understand how proper functions can be represented in those models. I begin in section 4.2 by introducing interventionism and making a general point: relations have an explanatory role to play in causal models. In section 4.3 I translate key technical terms from teleosemantic theory into the language of causal models. These terms are required for the definition of mapping relations. Finally, in section 4.4 I define mapping relations and outline their special explanatory role.

4.2 How relations explain

4.2.1 Motivation: worries about representational explanation

In the philosophy of cognitive science, there is a worry that representational explanation is just a gloss on underlying mechanistic explanations (Chemero, 2009; Hutto and Myin, 2013).[1] The thought is that mechanisms provide privileged explanations. On this view,

[1] It is not clear to me whether philosophers closely associated with contemporary accounts of mechanistic explanation, such as Machamer et al. (2000) and Craver and Bechtel (2007), would agree with my arguments in this section. The ecumenical spirit of the following quote gives reason to think that they might: "There may be non-mechanistic forms of explanation (for example, mathematical, geometrical, or intentional) that have different norms, and for which talk of mechanistic detail may be wholly inappropriate." (Craver and Kaplan,

describing inner states as *standing for* external states is not a literal scientific posit but convenient shorthand for an overall causal picture that characterises agential behaviour solely in terms of causal inputs, transformations, and outputs. A similar threat to representationalism is posed by the alternative dynamical systems approach (Gelder, 1998; Meyer, 2018). Instead of appealing to underlying causal interactions between objects and properties, dynamical explanations draw on equations that describe how properties of systems (and sometimes also relations between aspects of systems) change over time. For example, the Haken-Kelso-Bunz (HKB) model provides a dynamical explanation of human finger-wagging behaviour (Meyer, 2018). At the heart of the model is a differential equation[2] that describes how the *relative phase* of finger movement changes over time. Relative phase is a property of finger-wagging behaviour: hold up both hands and move your index fingers left and right; they will either both go left/right at the same time (relative phase of zero), or one will move left while the other moves right (relative phase of 180°). As the speed of finger-wagging increases, subjects spontaneously transition from in-phase to anti-phase movements. The dynamical equation at the heart of the HKB model describes this system, and predicts that spontaneous switching behaviour. Proponents of the dynamical systems approach to cognition hope that many other kinds of behaviour can be described in a similar way.

If either (or both) of the mechanistic and dynamical views are on track, the point applies to biological signals too. The best candidates for contentful states in natural science are cognitive states. If they lost their status as contentful, it is likely that simpler signalling sys-

2018, n.2 p.2).

[2] A differential equation describes how a property of a system changes with respect to another property. In dynamical systems models, the property with respect to which features of the system change is time.

tems would soon follow.[3] Explanations in terms of content would be considered shorthand for underlying mechanical and/or dynamical explanations. Focusing just on the mechanistic alternative, since appeal to mechanisms – objects, processes and properties – can do the requisite explanatory work, we seem to lose the incentive for positing representational content. And since there are metaphysical worries about content – especially given the possibility of falsity, cases where the content of a representation does not obtain – it would seem best to dispense with it.

In order to assuage this worry, teleosemantics defines content in terms of a certain kind of explanatory *relation*. The relation is borne between a signal and a world affair that influences the success of the signal's receiver. To spell this out, we must first look at how relational explanation works in general. How can an explanation of the behaviour of a causal system consistently and literally appeal to relations? One problem in answering this question is that there is no consensus theory of explanation (Woodward, 2019, §7.3). Fortunately, one contemporary approach enjoying real success is **causal modelling**, spearheaded by Pearl (2000) and Woodward (2003). Woodward's account of explanation suits our problem well, because we have already chosen to represent signalling interactions using causal models. This section introduces his account and applies it to a certain kind of relation-involving explanation. The following sections apply it to the teleosemantic theory of representation.

[3] It is admittedly conceivable that someone might claim biological signals are more securely contentful than cognitive states. One could argue teleosemantics is appropriate in the biological case because senders, receivers and proper functions can be readily identified, whereas cognitive/neural states are not so clear cut. Rosa Cao (2012, 2014) argues that the neural case is trickier than one might think.

4.2.2 Interventionism

Interventionism is an account of causal explanation (Woodward, 2003). It gives conditions under which an object or process X can be said to cause, or figure in a causal explanation of, an object or process Y. Those conditions are expressed in terms of the concept of an **intervention**. An intervention is a hypothetical experiment whereby a scientist 'takes control' of X, setting its value and breaking upstream causal connections. An intervention on X is written $\hat{X}$.

Roughly, interventionism interprets 'X is a cause of Y' as 'interventions on X systematically change Y, given certain background conditions N'. Suppose for example X is the temperature of a container of water, Y is the phase of the water (solid/liquid/gas), and N is the pressure inside the container. For normal atmospheric pressure, the container's temperature can be a cause of the water's becoming a gas. Intervening on X, increasing it beyond 100 degrees Celsius (by heating the container, for instance), would lead to the value of Y being 'gas'. Setting X to between 0 and 100 Celsius would lead to Y being a different value, 'liquid', while setting it even lower would lead to the value 'solid'. So the first part of the interventionist interpretation of 'X causes Y', that interventions on X systematically change Y, holds. The second part adds the caveat that the relationship need only hold under certain background conditions. This allows for the fact that the relationship between X and Y may be distorted, or disappear altogether, under different values of N. For example, different pressures alter the way in which temperature affects water phase. At low enough pressure, it is not possible to set Y to 'liquid', no matter what the value of X.

Background conditions affect causal relationships. By constraining causal explanations to within certain ranges of values those background conditions (N) can take, interventionism

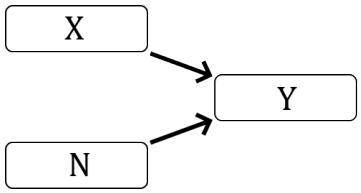

Figure 4.1: A simple model of the basic interventionist account. The account states that 'X is a cause of Y' means 'interventions on X systematically change Y, given certain background conditions N'. Example: X is the temperature of a container of water, Y is the water's phase (solid, liquid or gas), and N is the pressure. Since, given certain pressures, intervening on the container's temperature changes the water's phase, we can cite the temperature in a causal explanation of the phase.

mandates the appeal to upstream variables (X) in explaining causal effects (Y). Figure 4.1 depicts a basic relationship between cause, effect, and background condition.

The interventionist analysis allows explanations that describe certain *values* of X as causes of certain *values* of Y, as well as explanations that simply describe X as a cause of Y. As an example of the first kind of explanation, the container's temperature being above 100 degrees Celsius is a cause of the water's becoming a gas. As an example of the second kind of explanation, the container's temperature is a cause of the water's phase. Although Woodward's approach emphasises the second – variables causing variables – it will sometimes be useful to speak in terms of the first – values causing values.

Interesting properties of causal relationships can be investigated using the mathematical toolkit this framework enables. For example, Griffiths et al. (2015) argue that $I(\hat{X}, Y)$, the mutual information between interventions on X and Y, is a good measure of *causal specificity*, the extent to which X is a specific cause of Y. Intuitively, $I(\hat{X}, Y)$ quantifies how much control an intervention on X affords over Y. Greater control is a reasonable way to capture the informal concept of causal specificity. In the example of water, extremely low pressures

reduce the amount of control X has over Y by reducing the number of different values Y can take. This in general reduces $I(\hat{X},Y)$, and seems to capture the intuition that the hypothetical scientist has less 'control' over the phase of water because they are no longer able to make it become liquid. Not all ways of reducing control over Y involve removing one of its states from the space of possible values. But that is one particularly vivid way to illustrate the idea.

Let us see how relations can do explanatory work by adopting this framework.[4]

4.2.3 Relational explanation

What do relational explanations look like? Consider a machine that randomly selects two socks from the washing pile and pairs them *iff* they match. Otherwise they are returned to the pile, which is then stirred. The system can be modelled as per figure 4.2.

Suppose the machine picks two matching socks. It will pair them, and we can explain why they end up paired by saying that sock 1 matches sock 2. If we take the verb *matches* to describe a matching relation, we can express that relation as per figure 4.2. If instead the machine returns the socks to the pile, the corresponding explanation for this outcome may cite the lack of a match.

What is useful about this explanation is that it does not need to cite individual properties of socks. In particular, it need not state the pattern on each sock individually before deducing that they match. A more fine grained explanation would state the pattern on each sock before comparing them and pronouncing them a match. Descriptions like this are perhaps suited to

[4]Other explanatory frameworks may well have an easier time encompassing relational explanation. For example, the statistical relevance approach assigns explanatory relevance on the basis of conditionally dependent relationships (Woodward, 2019, §3). This does not require that the features whose dependence grounds explanation be one-place properties. If this is right, relational explanation is in even better shape than I argue for here.

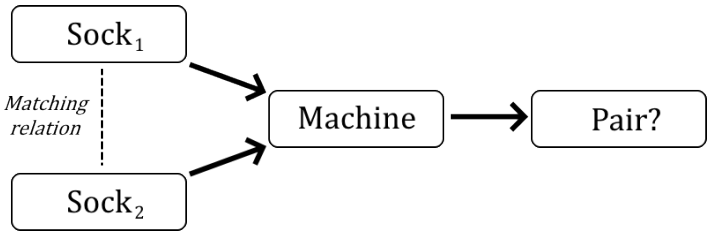

Figure 4.2: A causal model representing the sock-sorting machine. The variable **Pair?** takes value PAIRED if the socks match, and REJECTED otherwise. Explaining why the variable **Pair?** takes a particular value on a given occasion, we can appeal to the presence or absence of a matching relation between the socks.

tracking the machine's internal workings, but they are not a necessary part of an acceptable explanation of the outcome. Relational explanations are useful *because* they abstract away from individual properties.

A much simpler but more powerful version of this concept is a logical AND gate (figure 4.3). Suppose **Out** reads YES after one operation. We can explain this by saying both $\mathbf{In}_1$ and $\mathbf{In}_2$ were YES. However, this looks like the fine-grained property explanation rather than the efficient relational one. AND gates do not accept two NOs, so the relation we appeal to can only be *both*-YES, which seems to be nothing but a restatement of two one-place properties. We get a little explanatory purchase, however, when **Out** reads NO: we can say that the inputs were not both yes, which abstracts away from the three specific circumstances in which this condition holds (NO-NO, YES-NO, NO-YES).

For a better example, consider the XNOR gate. It has the same causal model as other binary logic gates, but accepts both YES-YES and NO-NO. Explaining the operation of the XNOR gate is easier by appeal to the sameness of the inputs. *Logical sameness* is a relation that affords modest explanatory purchase, just as *Pattern sameness* does for the sock-sorting

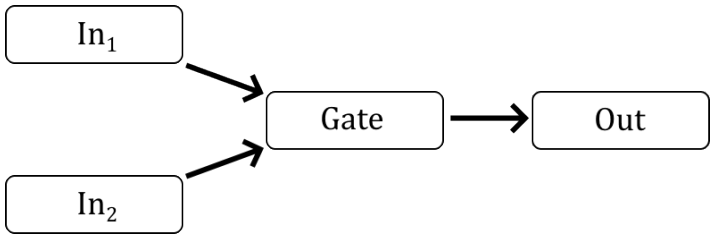

Figure 4.3: A causal model representing a binary logic gate.

machine.

Relations between nodes in causal models provide explanatory purchase. Logic gates are perhaps the simplest example. Their outputs are the result of relations between the inputs. XNOR does not care what its inputs are, only that they are the same. Abstracting away from properties to produce causal consequences as a result of relations is what makes logic gates useful.

A purist may object that relational explanation is at best shorthand, at worst misleading. Logic gates operate by comparing inputs *in a sense*, but describing them as such masks the whole truth. It makes it seem as though the relation itself were doing causal work. Since this cannot literally be the case, the gate must consist of an intricate causal structure which, appropriately hooked up to inputs and outputs, implements a logical operation. What happens inside the gate is causal and is explained solely by reference to objects and their properties. Consider, for example, the logical AND gate depicted in figure 4.4 as a domino chain. Its layout ensures the **Out** chain falls iff both **In** chains fall roughly simultaneously. There is no component within the gate that compares the inputs, there is only a causal structure that mimics such a comparison.

This objection comes in two flavours. First, the purist could be saying that relational

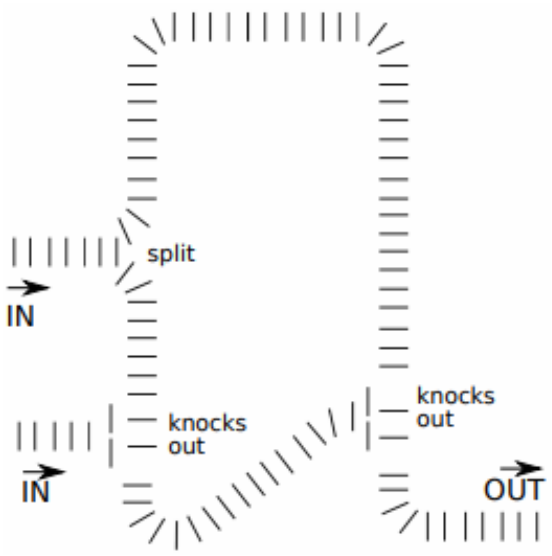

Figure 4.4: A logical AND-gate built using dominoes (from think-maths.co.uk). If neither in-chain is knocked down, the out-chain remains standing. If the first (uppermost) in-chain alone is knocked down, it interrupts itself and the out-chain remains standing. If the second in-chain alone is knocked down, it does not connect to the out-chain, which remains standing. If both in-chains are knocked down, the second interrupts the first's interruption, leaving the first chain free to knock down the out-chain.

explanation is a pseudo-explanation – it is (say) metaphorical or fictional rather than literal. Second, the purist could be saying that relational explanation is appropriate (literal) at one level, but that that level of explanation is not privileged. On this view, mechanistic (or dynamical) explanation has priority, perhaps because it has more predictive power, applies in a wider class of situations, or possesses some other desirable feature of explanation.

The first reading is more radical and the second is conciliatory but neither, I think, will stand. The first reading requires that objects and properties be real while relations are not; it is then better to dispense with talk of relations altogether in a completed science. But the examples of the sock-sorting machine and the logic gate show that we have no reason to treat relations as unreal. In these toy cases, relations are collections of, or abstractions from, pairs of one-place properties. This suggests that if properties are metaphysically secure, so are relations. To the extent we have reason to believe one-place properties are real, we have reason to believe relations are real too.

What the second reading requires is an account of privileged explanation, a reason to believe mechanical explanations have elite status. But it is just not clear why mechanical explanations should be universally privileged. Indeed, there is plenty of scepticism about the idea that *any* single kind of scientific explanation, whether mechanistic or otherwise, is going to be universally privileged (Woodward, 2019, §7.3) (Woody, 2015). Woody, for example, has recently argued that "the explanations that are generated and endorsed across modern scientific communities are diverse and pluralistic, rather than homogeneous, in kind" (Woody, 2015, p. 80). There are certainly contexts in which mechanical explanation is more valuable and fiercely sought. Whole disciplines – certain areas of physics and chemistry, for example – perhaps employ relational explanation as a shorthand or stop-gap until the

mechanical story is discovered. But the idea that scientific explanation *tout court* is beholden to mechanism is grossly unmotivated.

Admittedly, mechanistic analysis sometimes offers further benefits to the scientist. Identifying mechanisms helps scientists do two things that are more difficult in their absence: prediction and control. However, knowing a mechanism is not always necessary for prediction or control. Having observed my sock-sorter over a period of operation, I can predict its behaviour upon being presented with a new pair of socks. I can also control its behaviour: supposing I want it to always pair socks, I can ensure that any time a sock is placed in the first slot, the sock placed in the second slot matches it.[5] These abilities are less secure when the mechanism is unknown. There may be some principled limit on the accuracy of the machine's sock-scanner, such that two non-matching socks are mistakenly paired if they are similar enough. Knowing these limits would improve both prediction and control, and knowing the mechanism is one way to know the limits. Nonetheless, explanation, prediction and control can still go on without knowledge of the mechanism.

4.2.4 Relational explanation and causal role

I am contrasting mechanical with relational explanation. The term 'causal' is still unassigned. It is typical to treat **causal** as synonymous with **mechanical**, but there is good reason not to follow that practice here. If we follow Woodward (2003), causal explanation ought to be understood by reference to causal models. I have argued that these models support relational

[5] To anticipate, there are at least some cases where knowing the relevant mapping relations aids in prediction and control of a sender-receiver system: von Frisch was able to predict the flights of honeybee workers in his observation hives after watching the waggle dance; had he possessed sophisticated fake-bee technology, he might have been able to control those flights. To do either of these things, he did *not* need to know the mechanism by which receivers decode the dance.

explanation too. I will continue to use **causal explanation** to mean explanation understood with respect to causal models, which (I claim) includes both mechanical and relational explanations.

But do relations have a **causal role**? Let me stipulate that in order to have a causal role, an object or property must have a position within a causal model such that intervening on it affects something downstream. In Woodward's account of causal explanation, interventions are acts carried out on a single node – a single variable. So only things associated with single nodes can have causal roles, on my understanding of the term.

One way to establish the claim that relations have explanatory roles in causal models would be to argue they do indeed have causal roles. The thought would be that sock-pairers and logic gates are the means by which relations gain causal efficacy. On this view, the relation *Pattern sameness* really does cause the socks to be paired, and the relation *Logical sameness* causes the output of the XNOR gate to be YES. Two recent articles, by Gładziejewski and Miłkowski (2017) and Meyer (2018), sharpen this thought by treating relations as variables that can be intervened on. I want to explore their suggestions, in part to contrast them with my own account, but also to show the various ways in which relational explanation can be supported.

Gładziejewski and Miłkowski (2017) argue that relations have causal roles ("causal relevance" in their terminology) via the claim that it is possible to intervene on relations. Rather than treating relations as features of multiple nodes, they propose treating a certain kind of structural resemblance relation as a *single* node X. Intervening on X changes a downstream node Y (which need not itself be a relation). The authors use this account to argue for the causal relevance of representational relations. They favour a structural resemblance account

of the representation relation, on which 'S-representations' represent in part by being structurally similar to their representational content:

> For similarity to cause success, interventions into the value of X (which corresponds to the degree of structural similarity between the representational vehicle and what it represents) should result in systematic changes in the value of Y (which corresponds to the degree of success of the mechanism that makes use of an S-representation in performing its mechanistic function or capacity).
>
> Gładziejewski and Miłkowski (2017, p. 343)

This scenario corresponds to the basic interventionist pattern depicted in figure 4.1, letting X be a relation of similarity between vehicle and content, Y degree of success of the agent harbouring the representation, and N any relevant background conditions (not mentioned in the quote).

Meyer (2018) makes a similar claim by appealing to relations of a different kind. For him, the interventionist diagnostic for causal explanation is not restricted to causal models. Any model in which altering one variable leads to changes in another can be interpreted causally. Meyer is particularly concerned with dynamical models:

> If we can intervene on the values of variables in a dynamical model, and see changes in the value of another variable, then (on the interventionist account) we have exposed a causal relationship.
>
> Meyer (2018, p. 12)

Meyer draws on the HKB model as an example. That dynamical model incorporates the following differential equation:

$$\frac{d\phi}{dt} = -a\sin\phi - 2b\sin 2\phi \tag{4.1}$$

In equation (4.1), $\frac{b}{a}$ is a feature of the finger-wagging system known as the coupling ratio,

and ϕ is the relative phase introduced earlier. Both terms represent relations, and intervening on $\frac{b}{a}$ changes the value of ϕ. According to Meyer (2018, p. 2), applying the interventionist diagnostic "even-handedly" entails treating $\frac{b}{a}$ as a cause of ϕ.

My own view is weaker than those of Gładziejewski and Miłkowski (2017) and Meyer (2018). I propose we reserve nodes in causal models for objects and their properties. Relations therefore do not have causal roles. But they gain an explanatory role as a result of *patterns of invariance* in multiple interventions. Let me explain.

Combinations of interventions on causes reveal invariances in effects. Such invariances may be attributable to relations. For example, intervening on both inputs to the XNOR gate, setting them both to the same value (whether both YES or both NO), produces the same output, YES. A relation has an explanatory role when combinations of interventions that instantiate it produce the same outcome. Logic gates and sock-pairing machines are simple examples of a general phenomenon: relations are abstractions of collections of one-place properties, and the explanatory role of relations derives from invariances in causal affairs downstream of those collections.

It is not clear how much violence is done to the concept of an intervention. Traditionally, upstream variables other than the one being intervened on a considered background conditions, as per figure 4.1. I am proposing we consider the results of intervening on multiple nodes at once, in effect treating what would otherwise be considered a background condition as a potential cause. I do not think this invalidates my account. Woodward himself dedicates a chapter to a relevantly similar notion of invariance (Woodward, 2003, §6). He is concerned with invariance of generalisations, in the sense that they continue to hold under

certain interventions.[6] In any case, if Gładziejewski and Miłkowski (2017) can treat relations as nodes, and Meyer (2018) can appeal to interventions outside causal models, then appealing to patterns in the effects of multiple interventions is not overly radical.

The fact that there are patterns of invariance in sets of interventions underpins the explanatory purchase of relations. Appealing to a relation of logical sameness allows us to predict and explain the output of the XNOR gate without specifying the particular input values. Of course, the notion of explanatory purchase has not yet been rigorised. But it seems clear to me that if an appeal to relations allows you to ignore underlying properties and yet still make accurate predictions, there is a gain in explanatory purchase. For consider a gate that was able to produce four different outputs, and did so depending on the four different combinations of inputs. There could be no descriptive gain in appealing to logical sameness, because the two different conditions under which the inputs are the same lead to two different outputs. It seems that when there is a one-to-one correspondence between combinations of inputs and possible outputs, there is no point in grouping sets of inputs together. When many inputs produce the same output, however, that is a pattern of invariance; often, as in the case of the XNOR gate, there may be a relation or relations that explain it.[7]

Shea (2018, §2.3) provides similar justification for appealing to representational content in cognitive science. He accepts that descriptions of system behaviour can be given in purely

[6] One might tentatively suggest that Woodward's notion of invariance is related to my own as follows: the generalisation about how the sock-sorting machine works (matching socks lead to the **Pair** variable being PAIRED) is invariant (Woodward's sense) under interventions on either of the two inputs to the extent that **Pair** changes when the intervention changes the relation and does not change (is invariant in my sense) when the intervention does not change the relation. Spelling this out in further detail must wait for future work.

[7] Artificial neural networks operate on a similar principle. Edge detectors, for example, do not simply aggregate the one-place properties of pixels that constitute the edge. They exploit the spatial relations *between* pixels. That is what an edge *is*, in visual data: a relational configuration of pixels. Any system that can recognise the edge given access only to the pixels is dealing with relational features.

internalist terms; that is, descriptions that appeal only to causal properties. Nevertheless, there are significant categories of behaviour whose description can only be given by reference to external features of the environment. Shea claims most explananda in cognitive science are partly externalist. These include "explaining a system's reaction to and distal effects in its environment" (Shea, 2018, p. 231). The general class of abilities Shea demarcates that have this externalist feature he calls *task functions*. For our purposes, task functions share with proper functions the feature that distal properties and processes in the environment can affect their successful performance (see section 4.3). It is this feature, pervasive in both biology and cognitive science, that merits drawing on extrinsic properties in constructing the explanans. The account given of the explanatory role of mapping relations in section 4.4 is in broadly the same spirit.

A few final comments are in order before moving on. First, given the definition of causal role suggested above, we can characterise the less radical strand of purism as the claim that *causal roles are the privileged explanatory roles*. It is this claim I reject. Second, the account of Gładziejewski and Miłkowski (2017) supports my claim *a fortiori*. I claim that relations can play explanatory roles in causal models. I argued for the claim by outlining an explanatory role derived from patterns of invariance in interventions. However, a simpler way of playing an explanatory role in a causal model is just to have a causal role. And Gładziejewski and Miłkowski (2017) claim relations do have causal roles. We may disagree on the particulars, but we are at least unified against purism. Third, Meyer (2018) raises the possibility of applying interventionist criteria in non-causal models. Pursuing this line might lead to even stronger consilience between my own view and that of Gładziejewski and Miłkowski (2017): perhaps there is an appropriate kind of model in which relations appear

such that they can be intervened on. Currently we differ in opinion over whether causal models satisfy that need, but we might agree over another kind of model.

In sum, whether or not relations play a causal role, they play explanatory roles in causal models. Insofar as the causal modelling approach is a good account of explanation, there is an explanatory role for relations in natural science.

4.2.5 When are relations 'real'?

A sock sorter could be built that pairs any desired combination of socks. Suppose there are four possible sock patterns, Clouds, Flowers, Ducks and Fish. Assuming a one-to-one correspondence between left and right socks is required, there are 24 possible relations that a sock sorter could instantiate; left-foot socks with Clouds could be paired with right-foot socks with Ducks, and so on. Some of these unorthodox relations are listed in table 4.1. We tend to think, however, that the relation of *Pattern sameness* – hereafter **the matching relation** – is more 'real' or 'natural' than a relation artificially induced by an unusual machine. Why? On what grounds can we trust the intuition that the 23 'gruesome' sock-sorters do not respect a real or natural relation?

The intuition can be bolstered in two ways. First, it is easier to build a sock-sorter that respects the matching relation. If we say that relations are real to the extent they are exploitable, and that a measure of exploitability is the relative ease of building a machine that respects the relation, then the matching relation is more real than the others. Second, the matching relation is more readibly extensible. This is really a consequence of the first assertion. We will consider them in turn.

The matching relation is more readily exploitable. Suppose you were tasked with

Left sock	*Right sock, matching relation*	*Right sock, gruesome relation 1*	*Right sock, gruesome relation 2*	...
Clouds	Clouds	Ducks	Clouds	...
Flowers	Flowers	Fish	Fish	...
Ducks	Ducks	Clouds	Ducks	...
Fish	Fish	Flowers	Flowers	...

Table 4.1: Each right-sock column represents a possible sock-sorting machine. The standard machine is described by the first right-sock column: it matches socks with a cloud pattern to socks with a cloud pattern, socks with a flower pattern to socks with a flower pattern, and so on. This machine respects the matching relation between socks. Other machines could be built that pair socks in unusual ways. These do not seem to respect any 'real' or 'natural' relation; the relations they respect are gruesome or artificial (though some of the pairs may in fact match, as with gruesome relation 2). The main text discusses the intuitions behind the claim that the matching relation is more real or natural than the other possible relations.

building a gruesome machine. Socks are to be paired according to one of the 23 unnatural relations, two of which are displayed in table 4.1. How would you do it? We might imagine a scanner in each in-slot, feeding data into a sock-recognition program. The left-sock program processes the scanned image and outputs a string of 1s and 0s representing whichever pattern was detected; the right-sock program does the same. A central processor then takes the two strings and checks whether they conform to the desired relation (table 4.2).

The point is that the machine must implement an intermediate step of 'recognising' each pattern, translating it into its own inner language before engaging the lookup table to confirm the pair. One good thing about this system architecture is that the lookup table can be easily changed. The same machine can instantiate many different relations just by changing one small part of it and keeping all the rest (scanners, sock-recognition modules) the same. But the price of this flexibility is inefficiency. The machine must explicitly recognise each sock-

Left sock pattern / code	*Right sock pattern / code*
Clouds / 00	Ducks / 10
Flowers / 01	Fish / 11
Ducks / 10	Clouds / 00
Fish / 11	Flowers / 01

Table 4.2: A gruesome sock-sorter respects an unnatural relation by explicitly recognising each sock pattern, translating it into an inner code, and consulting a lookup table to determine whether the two socks 'match'. The system architecture is flexible but inefficient: flexible because different unnatural relations can be respected simply by swapping out the lookup table for another and keeping the rest of the system the same; inefficient because it requires explicit recognition of each pattern. The main text argues that the more natural matching relation (Clouds with Clouds, Flowers with Flowers etc) can be respected without the need for explicit recognition, and that this fact contributes to the intuition that this relation is indeed more 'real' or 'natural' than its gruesome alternatives.

pattern in order to instantiate any relation of which that pattern is a relatum. Its laborious explicitness goes hand-in-hand with its general-purpose scheme: it is *because* it derives binary strings in its left- and right-sock modules that its lookup table can be replaced.

Why do I say a machine that respects the matching relation is easier to build? For one thing, the machine need not explicitly 'recognise' a pattern, in the sense of translating it into an inner string, in order to test whether it matches another pattern. Here are two ways to test for a match without explicit recognition:

1. The system places the two socks together so they are physically contiguous, scans them as if they were one single piece of fabric, and engages a subroutine that attempts to distinguish them based on that image. The subroutine has an accuracy threshold: if it cannot reliably distinguish the location and orientation of each sock, that is probably because the patterns overlap in such a way as to make them indistinguishable. That can only happen when socks bear the same pattern. Clouds next to Ducks would be easy

to distinguish. Crucially, this does not require that the subroutine be able to recognise or identify any individual pattern. It need only tell when one pattern ends and another begins.

2. The system takes a picture of each sock and tries to digitally manipulate one picture to look like the other. Again, there is a threshold: if the number of computational steps required to transmute one image into the other is too large, that is an indication that the images are significantly different, and so the socks very likely do not match.

There are other ways to implement sameness-testing than those just described. The point does not depend on the details, except insofar as it is possible to test for sameness of pattern without explicitly recognising the pattern. (Explicit recognition is, of course, another way of doing it, but that would afford no improvement in efficiency over the gruesome machines.) Perhaps there will be scenarios in which explicit recognition is as easy as any viable means of sameness-testing. But I suspect that in the majority of cases sameness-testing is easier to implement, and that this fact is intimately related to our intuition that the matching relation is somehow more real or natural than the gruesome relations our machine could in principle respect.

The matching relation is more readily extensible. Suppose you buy a new pair of socks. If your machine matches socks without explicitly recognising their pattern, it is perhaps already capable of matching your new pair. Implementations of this kind have a certain inherent extensibility. As a result of the fact that they do not recognise individual patterns, they are able to function correctly when presented with novel patterns. In contrast, the machine that uses a lookup table relies on being able to convert scanned images of socks into proprietary binary strings. Buying a new pair would require reprogramming the machine (or

at least adding a new module). This all contributes to the intuition that your new socks bear the same relation to each other as each of your old pairs did – the matching relation. It is because the machine exploits this pattern that it is readily extensible to new manifestations of it.

A significant corollary is that the explicit-recognition machine does not exploit the matching relation even when its lookup table is organised so that it does only match same-patterned socks. It does not exploit the relation because it does not save any effort in its operation. It would be just as effortful to instantiate a gruesome relation.

Another way to see the same point is to ask a question that will become pertinent for mapping relations. Which interventions in the left-foot sock correspond to interventions in the right-foot sock for the purposes of preserving a certain outcome? For the lookup-table machine the answer is contained in the lookup table. Supposing the lookup table is as per table 4.2, setting the left sock to Ducks requires setting the right sock to Clouds in order to obtain the outcome PAIRED. Any intervention on the left sock has a corresponding intervention on the right sock that would preserve that outcome. In general, interventions on the left sock set it to a value described by a cell in the left-hand column of the table. The corresponding intervention on the right sock is picked out by the cell in the right-hand column of the same row. The outcome-preserving correspondence between interventions – call it an **intervention mapping** – is easily described: set the value of the right sock to the cell in the same row of the lookup table as the left sock's value. This specification does not uniquely pick out actual interventions. Every possible intervention mapping is picked out by this rule, because different lookup tables entail different outcome-preserving pairings. Of course, for a given lookup table, the rule uniquely specifies an intervention mapping. But

changing the relevant intervention mapping is as simple as changing the lookup table.

By contrast, there is a privileged intervention mapping in the case of the matching-relation machine. The intervention on one sock should be mapped to *the same intervention* on the other: setting the value of left sock to Clouds requires setting the value of right sock to Clouds, and so on. Why is this rule privileged? Because the machine is not capable of supporting any other. Talking of real or natural relations implies a kind of simplicity. The matching-relation machine is simpler than the explicit-recognition machine. Its simplicity constrains which intervention mappings could possibly track its behaviour.

What is more, the matching-relation machine's intervention mapping is extensible. The intervention that sets left sock to a novel value, say Pyramids, is matched by an equivalent intervention on the right sock. This is so even if this is the first time the machine has encountered a sock with a pyramid pattern. Intervention mappings that are specified by lookup tables are restricted by the finite, predetermined entries in those tables. If a new pair of socks is to enter into the intervention mapping, it must be explicitly introduced by means of extension (or replacement) of the lookup table. A lookup-table machine does not respect the matching relation because it does not support the outcome-preserving intervention mapping in the right way.

4.2.6 Summary

Relations have an explanatory role in causal models. Explanatory purchase increases to the extent that the same outcome is explained by multiple combinations of causal antecedents. These combinations of antecedents are abstractions from properties – they are relations. Relations are more real or natural to the extent that they can be more readily exploited, and to

the extent that this exploitation is more readily extended to new instances of the relation.

In order to apply this account to mapping relations, it will first be necessary to lay down the basic features of teleosemantics in the language of causal models. The next section introduces those features – centred on the concepts of *proper function* and *Normal explanation* – and suggests ways to understand them in terms of the interventionist framework.

4.3 Towards mapping relations: proper function in causal models

My goal in this section is to demonstrate how core features of teleosemantic theory (proper function, Normal explanation, Normal conditions, and so on) can be represented in causal models. The full definitions of these terms are given by Millikan (1984, §§1-2). It is not possible to capture every detail of each definition. Nonetheless, I aim to capture all details relevant for this chapter's central claim: that mapping relations have a secure explanatory role in natural science, where this explanatory role can be understood with regard to the interventionist framework. I will therefore introduce simplified causal models that capture the central features of each concept, building up to the characterisation of mapping relations in the next section. It will be instructive to see how relations of different kinds play explanatory roles with regard to proper functions, and to compare and contrast these roles with the purportedly special role of mapping relations.

4.3.1 Selection

A proper function is an outcome that an item was selected for bringing about. For example, the mammalian heart was selected for circulating oxygenated blood around the body. Past instances of hearts that in fact circulated oxygenated blood thereby contributed to the production of future hearts, through the inheritance of genes that contribute to the development of hearts. The selection process responsible for hearts is evolutionary and acts on genes. But selection processes need not be genetic. They need not act on genes over evolutionary time. What is required is just that there be recurrence of an object of a given type, and this recurrence is in part due to the causal effects of past instances of objects of that type. This abstract schema characterises trial-and-error learning and cultural selection too. The left-hand column of table 4.3 lists objects that have proper functions as a consequence of different kinds of selection.

It is not just objects that can possess proper functions, but also processes like behaviours. Processes have causal effects, and can be selected due to these effects. Again, different kinds of selection can give rise to processes with proper functions. The right-hand column of table 4.3 lists examples of processes that have proper functions as a consequence of selection.

Selection is the non-accidental differential recurrence of entities. Items of a type recur – are either reproduced or reconstructed, or brought about by some other means – because of a property or properties they possess. Items with proper functions are selected not (just) because of properties they bear, but because of causal effects they have.

	Object	*Process*
Genetic selection	Organs e.g. Mammalian heart	Genetically specified behaviours e.g. Waggle dance
Trial-and-error learning	Novel artefacts created accidentally and retained for their usefulness	Conditioned behaviours e.g. Pressing a lever to obtain food
Cultural selection	Socially learned artefacts e.g. Polynesian outrigger canoes	Socially learned behaviours e.g. Human natural language

Table 4.3: Examples of proper functions arising from selection processes of different types. Rows denote different selection processes. Columns denote different ontological categories. Many artefacts and behaviours may be subject to more than one kind of selection process.

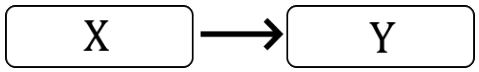

Figure 4.5: If X has as a proper function to bring about a particular value of Y, then Y must be causally downstream of X. In the simplest case, Y is immediately causally downstream of X. Example: a proper function of the mammalian heart X is to circulate blood Y. Values of X are different states the heart could be in, i.e. motionless, pumping, pumping erratically etc. Values of Y are different possible blood dynamics in the circulatory system i.e. static, circulating, circulating erratically.

4.3.2 Direct proper function

A **direct proper function** of an item X is a causal effect items of that type were selected to have. We will represent this as a value y_i of a variable Y causally downstream of X.[8] In causal models, causally connected items (variables) are represented as nodes with an arrow between them, as in figure 4.5.

In figure 4.5, X is a node that has causal effect Y. Y can take values y_1, y_2, ... and the value it takes depends only on the values of nodes that are causally upstream of it (in this case, X). Suppose X has as a proper function to bring about a particular value of Y, say y_2.

[8]It seems to me that the ontological type of a proper function is a *relation* between X and Y. For convenience, however, I will often speak of the relevant value of Y, here y_2, as being the proper function of X.

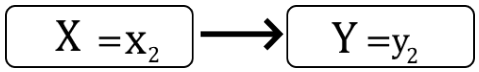

Figure 4.6: If successfully selected ancestors of X performed y_2 historically by going into state x_2, this forms part of the Normal explanation for how X performs y_2. Example: the heart circulates oxygenated blood (y_2) by going into the 'pumping state' x_2 (i.e. by pumping). Intervening on X and setting it to x_1, 'not pumping', would lead to oxygenated blood not circulating, y_1.

Then we can write $F(X) = y_2$. More verbosely, $F(X) =< 0,\ 1,\ 0,\ ...,0 >$ signifying that y_2, and not any of the other values of Y, is the proper function of X.

4.3.3 Normal explanation and Normal conditions

A **Normal explanation** says how members of the type X have performed y_2 historically. The appeal to history is constrained to just those instances that contributed to the selection of X in its present form. Normal explanations are causal explanations (or perhaps collections of causal explanations, one for each historical instance of an X causing a Y to take the value y_2). Normal explanation can therefore be captured in the interventionist framework. Suppose that in the past, when X took value x_2, Y took value y_2 (figure 4.6). We can consider interventions on each historical instance of X, changing it to (say) x_1, and ask whether the value of Y would have changed. Suppose that when X has any other value, Y has some value other than y_2. Then the way in which X historically performed y_2 is by going into state x_2. By establishing which changes in X would have led to changes in Y, we can establish how X historically performed y_2. This historical story is the Normal explanation for how X performs y_2. An equivalent locution says that X performs y_2 Normally by going into state x_2.

A more involved example results from adding interfering conditions, as per figure 4.7. Now X is partly at the mercy of N. Suppose $N = n_2$ is a necessary (but not sufficient)

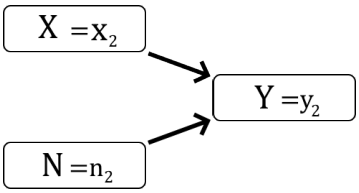

Figure 4.7: If N affects how X Normally performs y_2, then N must be mentioned as part of the Normal explanation. Example: N is the oxygenation of blood by the lungs, where n_1 = 'not oxygenated' and n_2 = 'oxygenated'. Then $Y = y_2$ only when $N = n_2$. Since this is a necessary condition, $N = n_2$ is (by definition) a Normal condition on the proper functioning of X: oxygen supplied by the lungs is a Normal condition on the heart circulating oxygenated blood around the body.

condition for Y taking the value y_2. Then $N = n_2$ would be part of the Normal explanation for how X performs y_2. Again, this relationship between the three variables can be understood on the interventionist framework (see figure 4.1 and the text explaining the relationship between cause, effect, and background condition). Conditions such as $N = n_2$ are **Normal conditions** for proper performance of y_2.

4.3.4 More or less proximate Normal explanations

Normal explanations can be more or less proximate. More proximate explanations mention only the causal connection between X and Y (and any intermediaries – see below) and the immediate conditions affecting Y (such as N). Less proximate explanations 'zoom out' and mention other relevant causal nodes. For example, in figure 4.8 the most proximate Normal explanation of how X performs Y mentions only that N necessarily takes value n_2. A less proximate explanation would mention which values of M ensure N takes that value.

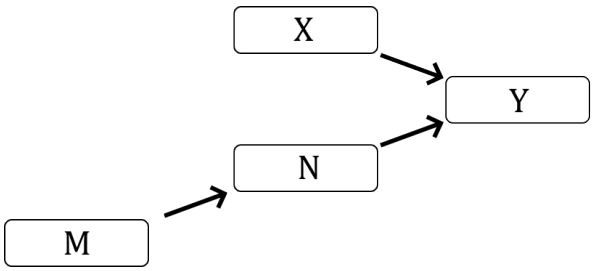

Figure 4.8: The most proximate Normal explanation of how X performs y_2 mentions only the relevant value(s) of N. A less proximate Normal explanation mentions which values of M produced those values of N. Example: the lungs (N) oxygenated blood by inhaling a gaseous mixture of oxygen and other elements from the air (M). An even less proximate explanation would mention parent nodes of M.

4.3.5 Intermediaries

A proper function need not be immediately causally downstream of the item to which it belongs. Consider figure 4.9. If X has y_2 as a proper function, Normal explanations of it will mention the values that A and B took when y_2 successfully occurred. Normal conditions include M and N, assuming their values mattered for the outcome y_2. For example, a heart X has the proper function to circulate oxygenated blood around the body $Y = y_2$. Intermediate stages include the various arteries A, B through which the blood travels, and the protection of those arteries by the skin M, N.

4.3.6 Relations as Normal conditions

It may well be that the clearest analysis of Normal performance mentions *relations* between causal participants. Consider figure 4.10. Suppose Y takes value y_2 when the value of A and N match (that is, when $A = a_1$ and $N = n_1$, *or* $A = a_2$ and $N = n_2$, and so on). Suppose further that Normally many of these different pairs occur, all leading to y_2. How best to

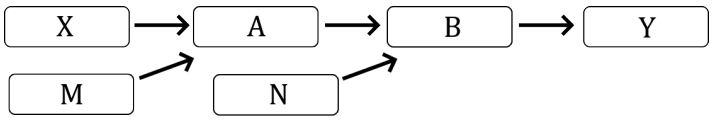

Figure 4.9: X must set the value of Y to y_2 by setting a value of A, that in turns sets a value of B, that sets a value of Y. Additionally, nodes M and N interfere with A and B respectively. The Normal explanation for proper performance of y_2 must mention the usual values of the intermediaries A and B, as well as the contributing values of M and N. Example: X is the heart, Y is the circulation of oxygenated blood, A and B are components of the arterial system through which blood flows, and M and N are external parts of the body (such as skin) that protect those arteries.

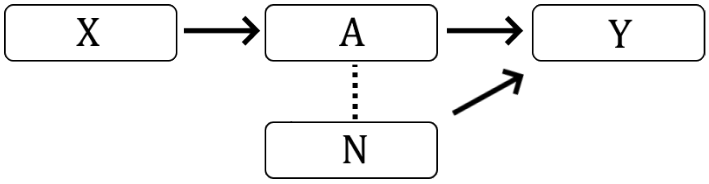

Figure 4.10: If the value of Y depends on a relationship between A and N, the Normal explanation of how X performs y_2 mentions this relation. Here the relation is illustrated by a dashed line, which is not part of traditional causal models. Example: let X be a foraging bee, A the set of possible locations it flies towards, N the set of possible locations of nectar, and Y the condition of obtaining or failing to obtain nectar.

explain how X works – what is its Normal explanation? It is not a disjunctive explanation that mentions each pair (a_1,n_1), (a_2,n_2),..., as if there were no pattern to these pairings. Rather, the Normal explanation mentions the relation between A and N.

Of course, the properties out in the world that are represented by values of variables like n_1, n_2 do not themselves come with indices like 1, 2, and so on. Numerical labels are indices of values of variables, which belong to models. Property values *themselves* do not come with labels. A 'match' between properties cannot be specified simply by saying that a_1 occurred and n_1 occurred, because the index subscript 1 is a human labelling convention.

In order for the Normal explanation to genuinely include a relation, that relation must be between properties in the world. There must be some reason why a correspondence between properties, represented by node values, participates in the Normal explanation. What reason? Let us first consider an example.

Here is a scenario that exemplifies figure 4.10 and includes as part of its explanation a genuine relation. A honeybee needs to forage nectar and bring it back to the hive. The bee (X) is causally upstream of two possible outcomes (Y): failing to find nectar (y_1) and finding nectar (y_2). As a result of a history of selection, y_2 is a proper function of X. The immediate activity that X can perform in order to bring this about is flying in a particular direction (A) from the hive. Finally, a Normal condition on proper performance is that there is nectar (N) in some direction from the hive.[9]

Under what circumstances will the bee be successful? It is not the case that flying in a particular direction will bring in nectar – it is not the case that flying in a direction represented by a particular value of A, say a_{17}, will achieve y_2. Nor is it just that there be nectar in a particular location (represented by a particular value of N). Rather, it is that *the location to which the bee flies is the same as a location of nectar*. Both A and N represent locations, and these locations must be identical for the bee to find and retrieve nectar as required.

Recall the analysis of relational explanation in section 4.2. When combinations of interventions yield an invariant outcome, that is a sign there is a relation at play. Intervening and setting the value of A, the bee's flight, to a certain location yields the outcome 'nectar obtained' Normally only when the value of N, the location of nectar, was set to the same location. The relation *Same location* is explanatory because there is an outcome-preserving

[9]For simplicity I ignore the condition that is the distance of nectar from the hive.

intervention mapping. That relation is part of a Normal explanation because the outcome that is preserved is successful performance of a proper function.

What is more, this location-matching relation seems 'natural' or 'real' for similar reasons to the pattern-matching relation in the case of the sock-sorting machine. Being in the same location as something else is easier to achieve than a gruesome relation between locations. Not only that, but it is just more likely that invariant outcomes will result from relations like this. The process that generates the invariant outcome 'nectar obtained' from the relative positions of bee and nectar is a simple process of physical transference of nectar from the flower to the bee. Relations that reliably lead to invariant outcomes as a result of purely natural processes – as opposed to artificial constructs like the lookup-table machine – seem more 'real' or 'natural' to us, as part of the same package of intuitions discussed in section 4.2. It would be extraordinarily unusual for an invariant outcome to be produced as a result of multifarious distinct causal antecedents, where no relation could be found between them. That would strike us as inexplicably fatalistic.

I am labouring the point about the naturalness of relations for reasons that will become clear in section 5.2. There we will face the question how Normal explanations can pick out the same relation across multiple historical performances of a proper function. For now, however, we must continue to put in place the theoretical devices out of which we will construct mapping relations.

4.3.7 Relations and cues

From the story so far, we can see how a Normal explanation might mention a relation between a causal antecedent of X and another Normal condition (figure 4.11). Suppose $Y = y_2$

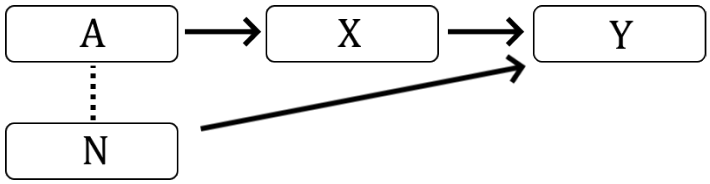

Figure 4.11: A relation between an upstream Normal condition A and a separate Normal condition N. This is one way to define cues. For example, if A is the location from which a sweet scent is emanating and N is the location of nectar, a bee (X) might condition its decision to fly towards a particular location (a node between X and Y, not depicted) on A in order to retrieve nectar ($Y = y_2$).

whenever (x_i, n_i) occur together. And suppose that X is designed to go into state x_i whenever $A = a_i$. If this is a successful means of setting $Y = y_2$, it can only be because (a_i, n_i) occur in tandem. The design of X is such that it conditions its behaviour on A *because* there is a relationship between A and N such that the value of A usually matches the value of N. That is part of what it means for A to be a cue.

The status of A as a cue signifying N requires the relation between them to be explanatory. As before, the interventionist framework allows us to determine whether the relation truly is explanatory, by examining combinations of interventions on A and N and their effect (via X) on the value of Y. Furthermore, the relation between X and N forms part of the Normal explanation too. We usually only consider the relation between the cue and the signified, but the device performing the function must itself be related to N in order to be successful.

4.3.8 Relational proper function

X has a **relational proper function** if it is supposed to "do or to produce something that bears a specific relation to something else" (Millikan, 1984, p. 39). Consider figure 4.12. Suppose X should set y_i when r_i obtains. Then X has a relational proper function.

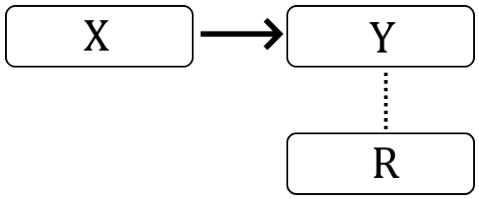

Figure 4.12: A relational proper function. X is supposed to produce a value of Y that bears a specific relation to the value of R. (Here I use R instead of N to differentiate between Normal conditions and relata that are part of the specification of a relational proper function.) Example: when a chameleon (X) needs to avoid detection, it is supposed to change its skin pigment (Y) to a colour identical to that of its immediate environment (R).

Recall that a proper function as originally described could be represented as a vector. For example, the function to produce y_2 could be represented as $< 0,1,0,0,....0 >$ where the vector positions are understood to range over the possible values of Y. Relational proper functions can be represented as matrices, signifying that proper performance is dependent on another variable. Equation (4.2) shows how a relational proper function of X, $F_R(X)$, can be represented as a matrix. Each column represents a value of Y, while each row represents a value of R. The presence of a 1 in a given cell signifies which value of Y must be set (the cell's column), given that the corresponding value of R occurs (the cell's row).[10]

$$F_R(X) = \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{pmatrix} \tag{4.2}$$

For example, a chameleon that needs to avoid detection should change its skin pigment

[10]If there were multiple relata with respect to which X should act, its relational proper function could be represented with higher-order arrays such as tensors. In what follows I assume for simplicity all relational proper functions are specified with respect to a single relatum, so can be represented by matrices.

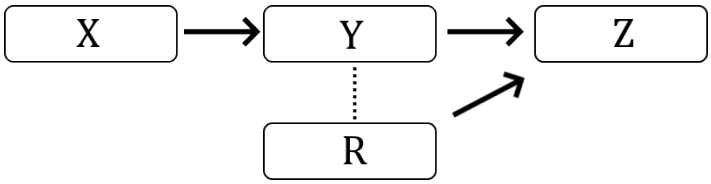

Figure 4.13: *Relational proper function: Y* is a relational proper function of *X*. There must be a downstream variable, *Z*, whose value depends on both *Y* and *R* and which contributes to the proliferation of *X*. *Adapted proper function:* Once a value of *R* is instantiated, the corresponding value of *Y* is an adapted proper function of *X*. *Derived proper function: Z* is a proper function of *X*, and a derived proper function of *Y*. It is an *invariant* (i.e. not *adapted*) function of *Y*. Example: *Y* is the skin colour of a camouflaging chameleon, *R* is the background colour, and *Z* is the outcome of being detected or avoiding detection by a predator.

Y to a colour matching that of the environment *R*. The rows of (4.2) correspond to different possible colours of the environment, and the columns to different possible colours of the chameleon. A value of 1 in a cell indicates that in order to perform its proper function, the chameleon must turn that colour (the column) when the environment is that colour (the row).

How can an item obtain a relational proper function? There must be some downstream result of both the effect *Y* and the relatum *R* that contributes to the proliferation of *X*. In the case of chameleonic camouflage, the effect is avoiding detection by predators. Figure 4.13 depicts this further variable as *Z*. As before, the interventionist framework can be deployed to ensure that it is in fact the relation that leads to the required outcome, and thus that *Y*, considered as a proper function of *X*, is indeed a relational proper function.

4.3.9 Adapted proper function

Suppose *X* has a relational proper function, like producing a pigment of a colour *Y* that matches the colour of its background *R*. This relational function is specified generically,

across the range of values Y and R can take. That is, the matrix given in equation (4.2) ranges over every relevant value of R, and specifies the appropriate value of Y for each. Suppose now that on a given occasion R has taken the value r_3. Then X acquires the **adapted proper function**[11] to produce y_3 (Millikan, 1984, p. 40). If the surface on which a chameleon is sat is green (r_3), then in order to perform its camouflaging function it must cause its pigment to turn green (y_3).

Relational proper functions are schematic, and can be represented by matrices. Adapted proper functions are specific, and can be represented by vectors. The appropriate vector for an adapted proper function corresponds to the appropriate row of the matrix representing the relational proper function from which it was bestowed. The adapted proper function corresponding to $F_R(X)$ with $R = r_3$ can be represented as a vector, equivalent to the third row of the matrix in equation 4.2:

$$F_{r_3}(X) =< 0,0,1,0,...0 >$$

The result of X's performance, the actual y_i that results, is an **adapted device**. The green pigment of a camouflaged chameleon is an adapted device.

In the language of interventionism, an adapted proper function results from treating the relatum of a relational proper function as a background condition whose value has been fixed. Holding fixed the value of R is tantamount to selecting a row of the matrix (4.2), defining the functional specification of X with respect to that value of R.[12]

[11] The word 'adapted' in the terms 'adapted proper function' and 'adapted device' (introduced momentarily) is not to be confused with the evolutionary sense of adaptation. Here, what the function is adapted to is the relatum R.

[12] A note on my chosen terminology here: although R is a background condition with respect to the causal

4.3.10 Derived proper function

Adapted devices have proper functions. Those functions are, more or less, what their producer produced them for doing. Since these proper functions are derived from the functions of producers, they are called **derived proper functions** (Millikan, 1984, p. 41).

Sometimes the further function, the ultimate job for which the device was produced, is a function of the producer that is always the same. Camouflaged skin colour is an example. Consider again figure 4.13 as a representation of the chameleon's predator-avoidance camouflage. Suppose z_1 represents being detected and z_2 represents avoiding detection. Avoiding detection is one of the functions of pigment production, so $Z = z_2$ is a proper function of X. Avoiding detection is also a function of the specific colour of pigment produced on a given occasion. Avoiding detection is not in itself an *adapted* proper function: it is always the same outcome that is needed in order to be successful (i.e. the outcome 'avoiding detection', represented by z_2). Therefore $Z = z_2$ is an **invariant derived proper function** of the pigment colour Y.

On the other hand, sometimes the further effect Z for which a device such as Y is produced is itself an adapted function. Then the derived function of Y – the function it has as a result of being produced by X – is an **adapted derived proper function**. For example, if Y is a honeybee waggle dance, its immediate function is to get watching bees flying in the direction of nectar. Letting Z be the flight path of the watching bees, both this and the dance are supposed to bear a relation to the **adaptor** R, the location of nectar. Figure 4.14 depicts

relationship between X and Z (because R influences the value of Z), the same is not quite true of the causal relationship between X and Y, because R is not directly causally connected to Y. Since the adapted proper function in question is a value of Y, it is not quite correct to say that R is a background condition for the causal relationship involved in this function. However, I can think of no better term, and I trust that this laxity will not result in confusion.

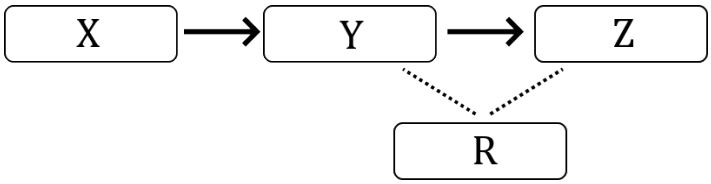

Figure 4.14: Z is a proper function of X, and a derived proper function of Y. Z is also an adapted function of X, so it is an *adapted derived* function of Y. Example: Y is a honeybee waggle dance, and Z is the flight path of watching bees. (It is not a requirement that Y be a signal, but examples involving signals are relatively easy to grasp.) In these kinds of cases there will be a further downstream variable (not here pictured) that is causally affected by both R and Z, e.g. the variable that represents finding or failing to find nectar.

this situation.

Whether Z is an invariant or an adapted derived proper function depends on whether the nodes downstream of *it* require there to be a relation between Z and R. Again, this can be revealed by an interventionist analysis.

In sum, the theoretical machinery in terms of which mapping relations shall be analysed are well-suited to being represented in causal models and understood with respect to the interventionist framework.

4.4 Mapping relations

4.4.1 The basic teleosemantic model

We now have all the ingredients of the basic teleosemantic model. The sender-receiver configuration depicted in figure 4.15 contains multiple functions of different kinds, as well as relations. To begin to analyse signals, we can investigate the basic model from the perspective of the *most proximate Normal explanations* for each of the functions of the Receiver,

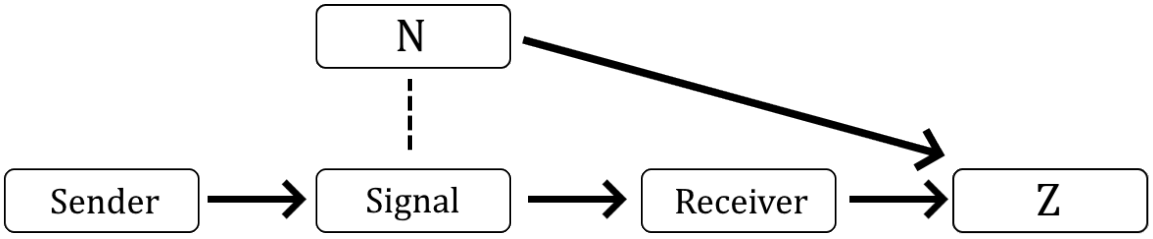

Figure 4.15: The basic teleosemantic model (figure 1.4, page 18), simplified and re-labeled. Several kinds of function and relation are present, though only the descriptive mapping relation is depicted.

Signal, and Sender respectively.

Receiver. In the simplest case, the receiver has a direct proper function. It is supposed to bring about a particular value of Z which is downstream of it. Z is also affected by N. If the receiver achieves the required value of Z by conditioning its behaviour on an upstream node, Signal, then Signal must bear a relation to N. This relation is a Normal condition on the receiver's functioning properly.

Compare this situation to figure 4.11 on page 164. There, an item achieved its proper function by conditioning its behaviour on an upstream variable. The upstream variable was characterised as a cue. Here, we are focusing on signals; however, by considering only the most proximate Normal explanation for the receiver's proper function, the analysis is the same as that given for cues. Since the most proximate explanation ignores the sender, the distinction between signals and cues collapses.

Signal. The signal has a direct proper function identical to that of the receiver. It is supposed to set a certain value of Z. Now, there is an intermediary between the signal and its proper function: the receiver. It is unusual to talk of the receiver as an intermediary, but from

the perspective of the signal, that is just what it is. The signal achieves its proper function by setting the receiver node to a certain value which, together with external condition N, produces the desired value of Z.

Since the value of Receiver must change depending on the value of N, Signal has a relational proper function to set the appropriate value of Receiver. In other words, the Signal performs its direct proper function (setting the value of Z, which should always be the same) by performing a relational proper function (setting the value of Receiver, which should differ depending on N). This situation is like the chameleon modifying its pigment colour to match the environment (N) in order to avoid detection ($Z = z_2$). On a given occasion, when the value of N has been determined, the Signal's relational function becomes an adapted function to set Receiver to a specific value.

As a consequence, Receiver is itself an adapted device, shaped by Signal into a form that helps it (the Signal) perform its proper function. For example, the waggle dance causes receiver bees to fly off in search of nectar. From the perspective of the dance, the flight is a means to an end; it is adapted by the dance to a form bearing the appropriate relation to the location of nectar. Normally, the dance causes a flight in the right direction by first bearing its own relation to nectar. The honeybee receiver's observation of, and reaction to, the dance is a sequence of causal mechanisms that convert the one relation into the other.

So far, we have two causally connected devices, Signal and Receiver. Both have identical direct functions – a value of the variable Z – and both perform their functions Normally by bearing a relation to a background condition N. Signal performs its direct function by first performing a relational function, configuring Receiver so it is capable of acting as required. It does this by bearing a relation to N. That same proposition, from the perspective of Receiver,

construes Signal's relation to N as a Normal condition. The Receiver uses this relation to perform the shared direct proper function, configuring itself so as to be appropriately related to N in the process.

The lesson to take so far is that when two devices share a direct proper function, its performance can be described from the perspective of either. Normal conditions – including relations – may take different roles in these two different stories. But the roles are equivalent so long as the devices cooperate.

Sender. As we are dealing only with cooperative systems, Sender shares the same direct proper function as Signal and Receiver. It must bring about that same desired value of Z. Sender can achieve this only through the intermediaries Signal and Receiver. Since the external condition N impacts the value of Z, Sender must cause Signal (and thereby Receiver) to bear appropriate relations to N. From the Sender's perspective, both Signal and Receiver are adapted devices. Setting their values is a relational proper function of Sender; setting their values on a particular occasion (i.e. given a value of N) is an adapted proper function of it.

As adapted devices, Signal and Receiver both gain derived proper functions. The Receiver now has *three* proper functions: its own direct function, a derived function from Signal and another derived function from Sender. Because the system is cooperative, all these functions pull in the same direction.[13] What is a good outcome for one is good for all. The

[13]It is not quite true that functions always 'pull in the same direction' in cooperative systems. All three functions of the receiver can have different satisfaction conditions on a given occasion if the sender makes certain kinds of mistake. Suppose a honeybee scout returns from a nectar source that has since (unbeknownst to the sender) become depleted; it cannot help the receiver perform its direct function (getting nectar) by helping it perform its sender-derived function (going to the just-visited source). Furthermore, if the sender performs an incorrect dance (perhaps because it is disoriented), then the receiver's signal-derived function is different again: it is supposed to go in the direction indicated by the dance, which is neither the just-visited source nor (barring

Signal has two functions, one direct and one derived (from Sender). Again they have one and the same satisfaction condition. In order to perform these functions, Signal must bear a relation to *N*. The relation between Signal and *N* is part of a relational proper function, from the Sender's perspective. It is a Normal condition for a direct proper function, from the Receiver's perspective. In other words, the Sender has as a proper function to ensure that a certain Normal condition exists for a proper function of a cooperating, causally downstream device. Part of what is special about mapping relations is this dual character.

The Sender too must bear a relation to *N*. In producing a value of Signal that bears a relation to *N*, the Sender itself must go into a state that bears such a relation. This need not be the same relation the signal bears. For example, bees that perform the waggle dance must bear a relation such as *just-returned-from-location-of* to the source of nectar their waggle dance is supposed to indicate. Otherwise, they do not perform their proper function (getting receiver bees to that same source) Normally. They might accidentally get receiver bees to a nectar source; that is a different matter.

Now we have a puzzle: if all three participants in the basic model – Sender, Signal and Receiver – must bear a relation to *N* in order for the system to work Normally, what is special about the relation borne by Signal?

4.4.2 What makes mapping relations special?

In the basic teleosemantic model, senders, signals and receivers all bear relations to *N*. But it is the relation borne by signals that is singled out for special treatment. The signal-signified relation is thought to be the fundamental representational relation, capable of playing a

lucky accidents) any nectar source at all.

unique and long-sought after explanatory role in biology and cognitive science. Why?

The question brings two desiderata along with it. First, in section 1.5 I claimed that paradigmatic signals are those for which sender and receiver are cooperating. The above analysis of the basic teleosemantic model assumes that all three components act to achieve the same end. This underpins the several layers of functions possessed by the signal and receiver, as well as the dual character of the relations borne by all (each relation is a Normal condition for more than one proper function, even though all pull in the same direction). If the system did not cooperate, these relations would not coincide. For example, the Receiver would require that the Signal bear a certain relation, while the Sender would require that it bear a different one. The special explanatory role of mapping relations should only be available when systems are cooperative. Or to put the point in less absolute terms, mapping relations should be explanatory to the extent that systems cooperate.

The second desideratum has to do with *resource separation*, the other paradigmatic feature of signals introduced in section 1.5. Paradigmatic signals are those for which the resources required to achieve the receiver's proper function are provided from elsewhere than the signal. Good signal design respects resource separation. The energy and materials expended in signalling are distinct from the energy and materials expended in achieving whatever outcome is guided by the signal. The specialness of mapping relations, then, had better respect this feature of the paradigm. It had better be the case that the unique explanatory role of signals and their mapping relations comes into sharper focus as resource separation increases. Otherwise we would lose the motivation to treat resource separation as paradigmatic.

As it turns out, these two desiderata are directly involved in accounting for the special

explanatory role of signals. When the basic teleosemantic model holds, three things can be said:

1. Greater cooperation enables greater resource separation.
2. Greater resource separation entails the signal's bearing the relation is more explanatorily important.
3. Mapping relations are special to the extent that they are the only explanatory feature of the signal vehicle.

I will discuss each statement in turn.

Greater cooperation enables greater resource separation. The basic teleosemantic model includes principles of good signal design. If a sender and receiver desire the same outcome, and the receiver has their own energy source, it is not worth attempting to transmit energy across the causal link between sender and receiver. Energy can be easily dissipated, and the system is already expending resources in sending the signal itself. It is more efficient to include only those materials required to overcome noise, allowing the receiver to accurately observe the signal. But this is only so when receiver and sender cooperate. Otherwise, if the sender has a function the receiver does not share, the sender may attempt to physically force the receiver into a certain course of action. The greater the antagonism, the more energy the sender must invest in forcing the receiver's hand. Conversely, the greater the cooperation, the more the sender can concentrate on getting the signal observed. This is just what resource separation entails.

Greater resource separation entails the signal's bearing the relation is more explanatorily important. When the signal's energy and material resources are *not* transferred to the receiver, the explanatory link between them must be relational. In the simplest case, the

signal triggers a response from the receiver, timed to coincide with a relevant external condition. Then the relation borne by the signal to the condition is 'at the same time as'. By triggering an immediate response, the signal 'transfers' this relation to the receiver. In other cases, whatever kind of relation the signal bears to the world, it is as a result of this that the receiver is configured (in Millikan's terms, adapted) to act appropriately. To the extent that energy and materials are not passed between signal and receiver, the only explanatorily relevant feature of the signal is the relation it bears.

Mapping relations are special to the extent that they are the *only* explanatory feature of the signal vehicle. Signals are special because their only job is to bear a relation to something else. It is this feature that prompts informational and representational descriptions from biologists and cognitive scientists. The relations borne by senders and receivers are overshadowed by their mechanical operation.

A similar feature of signals underpinning their special explanatory role has been recently noted by Artiga (2020). Paradigmatic signals do not provide material resources or parts of mechanisms that *enable* receivers to perform their responses. Signals are not **enabling causes**; they do not make it possible for receivers to do things they could not do otherwise. The acts receivers perform are those they could have performed without the signal, in principle. For example, receiver bees could in principle fly to any location within foraging range without observing the waggle dance. Signals *cause* receivers' actual flights but they do not enable them. Artiga (2020, §4.2) gives a fuller account of the notion of an enabling cause, and defines by contrast a **minimal cause**. He argues that paradigmatic signals are minimal causes: they cause receiver responses without enabling them. If signals are minimal causes, their explanatory role is purely relational. This further underpins their special status.

By contrast, the causal roles of senders and receivers in the basic teleosemantic model are not minimal. Consider first senders, and how they produce the three kinds of signal discussed in section 1.2: physical emissions, morphological structures, and behaviour. Physical emissions cannot bring themselves into existence and transmit themselves to receivers without the activity of their producers. And it is barely coherent to think of structures and behaviours existing without their bearers. In each case, senders are enabling causes of signals. Consider now receivers. Most kinds of response usually considered are behaviours. For example, foraging and predator avoidance are typical cases of inter-organismal signalling functions. The receiver is an enabling cause of these activities by the same argument as for the sender. My supposition is that in all paradigmatic applications of the basic teleosemantic model, receivers are enabling causes of their actions.[14] Senders and receivers are both enabling causes while signals are minimal causes. This helps to capture the uniquely relational explanatory role of signals.

4.5 Conclusion

Mapping relations have an explanatory role to play in the functional sciences. Evolutionary biology and cognitive science treat signals as contentful, and cite signal content in explaining behaviour. These practices are well interpreted by teleosemantics. According to the theory, Normal explanations of proper functional performance can mention relations. This approach is supported by an understanding of relational explanation in the interventionist framework. Causal models of proper functions reveal many kinds of relation borne by differ-

[14]The 'Act' variable was omitted from the simplified teleosemantic model in figure 4.15. It stands between Receiver and *Z*.

ent kinds of mechanism. The special explanatory role of mapping relations is secured by the unique causal profile of signals. Paradigmatic signals are minimal causes, designed so that the energy and materials used to transmit them are not used by the receiver in performing the response. It is this resource separation that ensures the explanatory role of signals is relational.

Chapter 5

Further issues with mapping relations

5.1 Introduction

The previous chapter introduced mapping relations and outlined their explanatory role. This chapter goes into more depth, responding to some objections to Millikan's theory and drawing links between it and communication theory. In section 5.2 I consider a proposed amendment to the theory due to Shea (2013). In section 5.3, I respond to an objection due to Godfrey-Smith (1996, §6) and Shea (2007). In section 5.4 I consider the relationship between mapping relations and communication-theoretic codes, as well as the question how mapping relations can play the explanatory role philosophers usually attribute to semantic content. In section 5.5 I argue that an important dual relationship in communication theory, between compression and transmission, can be understood in terms of the two fundamental kinds of mapping relations: descriptive and directive. Finally, in section 5.6 I offer an interpretation of the explanatory pattern that treats semantic content as a (one-place) property of signals, rather than as a relation.

5.2 Ontology of mapping

Two questions about the ontology of mapping relations have not yet been answered:

1. What exactly are we appealing to when we cite a relation in a Normal explanation?

2. If mapping relations are defined in terms of success, how can they explain success?

The second question is discussed in section 5.3. The reason the first has not yet been answered is that we have not fully considered the historical nature of Normal explanations.

5.2.1 The role of relations in Normal explanations

Normal explanations state how ancestors of a given (present) device performed their proper functions on those occasions in the past when they contributed to the proliferation of the lineage that gave rise to the (present) device in question. Normal explanations are aggregate explanations. They pick out what is the same across a range of historical instances.

The problem with citing a relation in a Normal explanation is that relations comprise many *different* token-token correspondences. Normal explanations are supposed to say what was the same across all these different instances. For example, a given honeybee dance has many ancestors, all pointing (let's suppose) in different directions. Yet the Normal explanation for how receiver bees successfully obtained nectar must say that all these dances were doing the same thing. What allows us to say that all these different dances bear the *same* relation to their nectar sources? What ensures the Normal explanation is univocal, as teleosemantics demands? We need a principled restriction on which historical instances count as falling under 'the same relation' for the purposes of participating in a univocal Normal explanation. In one sense, we are asking what allows us to 'abstract' from individual properties to a relation, where abstraction is as described in section 4.2.

Let us illustrate the problem in more detail before trying to solve it. Recall the gruesome sock-sorter from section 4.2. It pairs left-foot socks with a cloud pattern to right-foot socks with a flower pattern, left-foot socks with flowers to right foot socks with ducks, and so on,

with no paired socks ever having a matching pattern. I called the relation instantiated by this machine gruesome, but it has at least one important regularity: every time a left-foot cloud-pattern sock is placed into the machine with a right-foot flower-pattern sock, it will pair them. Each pairing will recur, and will recur *because* the machine has a lookup table which does not change during its operation. The pairings seem unusual, but at least they are consistent. Supposing this machine had as a proper function to produce these unusual pairings, a univocal Normal explanation could be given that cites its lookup table.[1] By contrast, imagine a machine whose lookup table is randomised after each pairing event. No consistent pairings would occur, so no univocal Normal explanation of this kind would be available.[2]

Relations cannot be cited in Normal explanations, explaining how a lineage of devices produced the same outcome by performing many different actions, unless there is some *reason* why those actions led to the same outcome each time. The aim of this section is to state conditions on there being such a reason. First let us survey what Shea and Millikan have said on the topic.

5.2.2 Shea's critique of Millikan

Shea (2013) claims that Millikan places a certain condition on citing relations in Normal explanations.[3] He claims she requires that there be an isomorphism between the two relata.

[1] In this case the Normal explanation might not be historical. The selection process that produces such artifacts is usually explicit design rather than a prolonged period of repeated selective events. For simplicity I will ignore this distinction. I assume the historical case is harder to solve, so by solving it we give good grounds to think the problem can be solved (if it even arises) for explicit design.

[2] Of course, if the machine was designed to randomise sock pairings then there would be a univocal Normal explanation. The explanation would cite its randomising program.

[3] The discussion between Shea and Millikan focuses on mapping relations. I want to emphasise that this issue concerns the role of any kind of relation in a Normal explanation.

In causal modelling terms, the condition is that there be an isomorphism between the values of the variables. The idea is that an isomorphism can provide the 'sameness' that unifies disparate instances. Different performances of a proper function, such as the chameleon's avoiding predation by matching its immediate environment, fall under the same Normal explanation because they all exemplify the same isomorphism – which for the chameleon is an isomorphism between the possible colours it can display and the possible colours of its immediate environment. This condition can be formulated as follows:

> ISO: For a relation to be cited in a univocal Normal explanation, it must be an isomorphism.

Shea claims that Millikan intends this condition to be an additional constraint on Normal explanation, beyond what has already been said in chapter 4.

As Shea points out, there is at least one significant problem with this proposal. The constraint that there be an isomorphism between two variables does not uniquely specify any particular relationship between them. Isomorphisms, construed as one-to-one mappings between the values of variables, are cheap. Any two variables with the same number of values will have many isomorphisms between them. In the case of the chameleon, the mere fact that a scientist can state an isomorphism between skin colours and environmental colours plays no special explanatory role. Isomorphisms therefore cannot help a relation to play a more substantive explanatory role in a Normal explanation. There are many different isomorphisms between skin colours and environmental colours that do not help explain how a proper function was performed on many historical occasions. In other words, since there would be isomorphisms in non-univocal cases, isomorphisms alone cannot constrain Normal explanations to be univocal.

Here it might be thought we can draw on the analysis in section 4.2. There I argued for a somewhat coherent package of intuitions that contributes to our treating relations as more or less 'real'. It might be thought that the more real a relation, the more univocal the explanation that adverts to it. There is something right about this idea, as we will see in a moment. But it is not the whole picture. Recall that the original 'gruesome' sock-sorter could perform a (relatively unusual) proper function in accordance with an acceptably univocal Normal explanation. What we are looking for this time is not a condition that distinguishes the standard sock-sorter from the gruesome one, but a condition that distinguishes both from the randomising machine. We can find it by following the discussion between Shea and Millikan a little further.

After rejecting the isomorphism constraint as too weak, Shea considers a stronger sense in which there can be a correspondence between the values of two variables. There can be **structure** within a variable, such that the values within it are related to each other. For example, the location of a nectar source is a variable whose values are all spatial locations. Spatial locations are related to each other spatially: the location in which I am currently standing bears the relation '30 metres north of' to a spot 30 metres south of here; that spot in turn bears that same relation to a spot even further south. The way mathematicians conceive of structure is as operations on, or **transformations between**, elements in a set. A spatial transformation on a location produces another location.[4] A transformation on colour produces another colour. There may be multiple kinds of transformation that relate the values of a variable, such as hue, saturation and luminosity for colours. However, only some of these may be

[4]Mathematicians would usually call a transformation a *function*, but I am strictly reserving use of that word for obvious reasons. Another way to think of a transformation is as an *operation* on one value of a variable that produces another value of the same variable.

relevant in a given case. Suppose for example chameleons were only capable of changing the hue, and not the saturation or luminosity, of their skin pigments. Then the transformation relating values of the variable 'chameleon skin colour' would be a transformation on hue, and would not affect saturation or luminosity.

How might structure ground the univocity of Normal explanation? The idea is that structure can relate different instances of performance of a proper function. Suppose for example we seek the Normal explanation for the chameleon's successfully avoiding predators. We look back in time, and find among the chameleon's ancestors a case in which one turned brown when its immediate environment was brown, and another turned green when its immediate environment was green. In order to give a univocal Normal explanation we have to show that the relation one ancestor bore to its environment is the same relation as that borne by the other. In one sense they are different relations – one is brown-brown while the other is green-green – but in another sense they are the same relation, namely *Same colour*. We appeal to structure: the colour transformation that changes brown environment to green environment is the same as the transformation that changes brown skin to green skin. The mapping brown-brown is therefore the same as the mapping green-green because this pair of mappings *preserves the structure* within each variable (figure 5.1). Preserving structure means that if you start with one value (brown skin), move to what it maps to (brown environment), then perform the transformation (green environment), you are in the same place as if you had first performed the transformation (brown skin to green skin) then followed the mapping (green skin to green environment). Situations in which skin did not match environment therefore do not belong within the Normal explanation. They do not obey this structure-preserving feature.

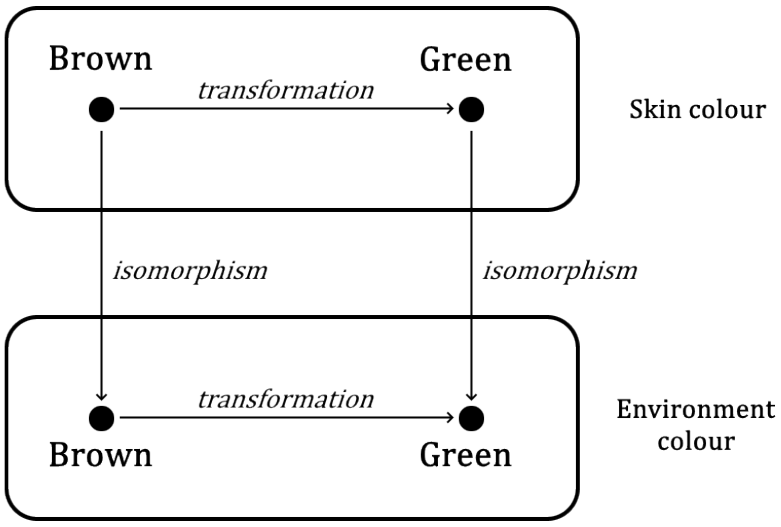

Figure 5.1: A structure-preserving isomorphism between two variables (which can be thought of as nodes Y and R from figure 4.13, page 166.) The isomorphism maps skin colour (upper variable) to environment colour (lower variable). The isomorphism is structure-preserving because performing the skin-colour transformation followed by the isomorphism yields the same result as performing the isomorphism followed by the equivalent environment-colour transformation. Two values of each variable are shown, to illustrate the point. Starting from the skin-colour value Brown, performing the brown-green transformation yields Green skin colour, and subsequently performing the isomorphism yields Green environment colour; starting again with Brown skin colour, performing the isomorphism first yields Brown environment colour, and performing the equivalent transformation produces Green environment colour. In mathematical terminology, the diagram commutes.

The proposal is therefore that isomorphisms *between* variables need to be supplemented with transformations *within* variables in order to enter into univocal Normal explanations. Isomorphisms that obey the structure defined by a set of transformations are called **structure-preserving isomorphisms**.

On the face of it, this proposal is promising. Structure seems to guarantee a tighter link between isomorphisms, motivating treating different token-token correspondences as part of the same relation. But Shea points out that at least one way of framing this condition is too weak, just as ISO was too weak. Shea considers this new constraint under the following guise:

> STRUC: For a relation to be cited in a univocal Normal explanation, it must be a structure-preserving isomorphism.

The problem with this condition is that it does not say which structure within the variables is to be preserved. For any gruesome isomorphism, there are gruesome transformations that count as grounding structure for that isomorphism. Structure, it turns out, is as cheap as isomorphism. For example, consider an isomorphism between skin colour and environment colour that takes brown skin to green environments and green skin to brown environments. This isomorphism does not preserve the structure depicted in figure 5.1, because performing the isomorphism followed by the environment-colour transformation depicted there would not yield the same result as performing the skin-colour transformation followed by the isomorphism.[5] However, the isomorphism does preserve a structure defined by a transformation that takes Green environments to Brown. This relationship between structures of skin colour

[5] In fact I did not specify a target for the transformation on Green environment-colour in figure 5.1. I think the general point should be clear nonetheless.

and environment colour is gruesome, because the first transforms Brown skin to Green skin (as per the original example) while the second transforms Green environment to Brown environment – and the isomorphism implies these are *equivalent* transformations.

Structure is cheap because there are gruesome transformations, and nothing we have yet said prohibits pairs of variables being associated with transformations that seem to us very unusual. What we need is a way to determine when transformations are genuinely equivalent to each other, in the way that *brown skin* $\rightarrow$ *green skin* seems to be equivalent to *brown environment* $\rightarrow$ *green environment*. Neither ISO nor STRUC can provide this.

5.2.3 The right way to constrain relations in Normal explanations

Neither ISO nor STRUC appropriately constrains the kinds of relation that can appear in Normal explanations. There are too many isomorphisms that satisfy each condition.

The problem with STRUC is that any given isomorphism preserves *some* kind of structure. The constraint could be tightened if we could specify what structure the isomorphism is required to preserve. That is, rather than saying 'there must be some pair of transformations that this isomorphism obeys' (which is too weak because there are plenty of gruesome transformations), we can say 'given *this* pair of transformations, there must be an isomorphism that preserves the structure'. This strengthens the condition considerably. It is much more demanding to preserve a particular, specified structure between variables than to satisfy STRUC. We can therefore examine the following condition:

> STRUC-FIRST Given a set of transformations on two variables, for a relation between those variables to be cited in a Normal explanation, it must be an isomorphism that preserves the structure defined by those transformations.

STRUC-FIRST is stronger than STRUC because it pre-specifies the structure that must be preserved, rather than leaving it an open choice (Shea, 2013, p. 67).

However, we now have the problem of determining which transformations are to be specified. We have a proper function for which we seek a univocal Normal explanation, one component of which is to be a relation. We can appropriately constrain that relation *if* we can specify the structure of each relatum. So the question is: how does a proper function pick out the structures to be preserved?

Millikan explicitly says that "Normal explanations are causal explanations" (Millikan, 2013c, p. 82). A proper function is a value of a node, and a Normal explanation is a causal explanation of how that value came about on many historical occasions. Relations are allowed to enter into Normal explanations in the same way that they are allowed to enter into causal explanations more generally. We can therefore draw on the account in section 4.2. If a Normal explanation is to mention a relation, it must specify *patterns of invariance* in interventions on the relata that bring about that value.

Interventions are more coarse-grained than transformations. Interventions specify a target value, while transformations specify both an original value and a target value. Nevertheless, discovering a pattern of invariance is equivalent to discovering structure of the kind depicted in figure 5.1. Take the case of the chameleon again. Having discovered a pattern of invariance, we can answer the question: 'Given the outcome of a brown skin-brown environment pairing, what is the colour of skin that produces this same outcome when the environment is green?' Interventionism delivers the answer 'green skin', completing the diagram in figure 5.1 in the required way.

Finding a pattern of invariance is equivalent to uncovering the required corresponding

structure of the two related nodes *and* the relation between them. So the demand that Normal explanations be univocal is already strong enough to constrain the relations that can appear in them. This demand implies STRUC-FIRST for relational explanation. Therefore, Millikan need not add an extra condition to the theory that ensures there exist isomorphisms for Normal explanations to cite. The condition that Normal explanations must be both causal and univocal already implies everything about isomorphisms that needs to be said. Shea (2013) appears to come to the same conclusion, though he does not draw on interventionism.

So far we have spoken of relations that appear in Normal explanation in general. What about mapping relations?

5.2.4 The structures preserved by mapping relations are determined by receiver design

Mapping relations are a special case of relations that enter into the explanation of relational proper functions. So mapping relations too must be isomorphisms that preserve structure, where that structure is determined by invariant outcomes preserved over many different occasions. What is special about mapping relations is that the invariance is determined by the way receivers respond to signals. Let us unpack this idea.

Relations appear in Normal explanations when downstream causal processes are sensitive to the structures of the relata. In the case of mapping relations, it is the receiver whose sensitivity to different signals brings about the invariant outcome. So the structure that must be preserved by the mapping relation is determined in part by the constitution of the receiver, by the differences it is causally sensitive to. This is the sense in which mapping rules are determined by receivers (which is, incidentally, what primarily distinguishes Millikan's

teleosemantics from that of Neander and Dretske).

Mapping relations have more scope to be unusual than relations appearing in relational functions in general. Receivers can be designed *along with* senders, and hence signals. Relational functions in general are hostage to causally downstream world affairs outside the scope of the selection process that shapes the device possessing the function. For the chameleon, the existence of predators is a given. It must match its skin colour to the environment, and the specification for the matching relation (what structure it must preserve and how precise it must be) is a consequence of the predator's perceptual system. Selection cannot tune both chameleon and predator to increase the efficiency wih which chameleons avoid predators. Indeed, in this case two selection processes work against each other. The point to bear in mind is that what is causally downstream of a relational function is in general outside the remit of the selection process that gives rise to the function. By contrast, mapping relations enjoy a certain kind of freedom that is a consequence of joint selection. Nonetheless, mapping relations must still be structure-preserving isomorphisms if they are to enter into Normal explanations.

However, here Shea's complaint seems to return with new force. Signals would bear an isomorphism to states – even a structure-preserving isomorphism – just by being of the same cardinality. Given a collection of success-relevant world affairs the receiver needs to act in relation to, *any* set of signals would bear *some* structure-preserving isomorphism to those states. Wouldn't such an isomorphism then count as a mapping rule? And wouldn't that make mapping relations to cheap to be of interest?

No, it is not the case that any specifiable isomorphism between world states and signals counts as a mapping rule. This is because the structure of the signal that the mapping rule

must preserve is determined by the constitution of the receiver. And the structure of the world states is constrained by the act the receiver must perform (just as in the basic relational proper function case). Thus both structures are somewhat constrained. The demand that there be a univocal Normal explanation implies STRUC-FIRST, as before.

It is true that if there were a completely free choice of receiver, then there would be a completely free choice of signal. Then it would indeed be easy to instantiate the appropriate isomorphism. But that is so rarely the case in real life, and even rarer in biological systems. There are constraints on what receivers are capable of, constraints on energetic resources, and constraints on what signals can be sent (that is, constraints determined by physical properties of the channel). The structure of the world states that must be preserved can*not* be freely chosen, for it is determined by the receiver's own relational proper function (see figure 5.2). Achieving an isomorphism that preserves that structure, given the constraints facing sender, signal and receiver design, is difficult. It is for this reason that isomorphisms that can be identified with mapping relations are not ubiquitous.

To sum up this section: Shea pointed out that if ISO or STRUC were intended by Millikan as extra constraints on which relations could enter into Normal explanations, they would be doing no good theoretical work. Shea argues (and we can agree) that the only condition of this kind that could do the relevant theoretical work would be STRUC-FIRST. Fortunately, STRUC-FIRST is entailed by the condition that Normal explanations must be univocal. Signals, as entities with relational proper functions, are unusual. They need to possess structure that, together with the structure of world states, is preserved by the isomorphism (the mapping rule). But there is a kind of freedom of choice in what structure they have. The arbitrariness of some signals is a consequence of them exercising this freedom. There is still

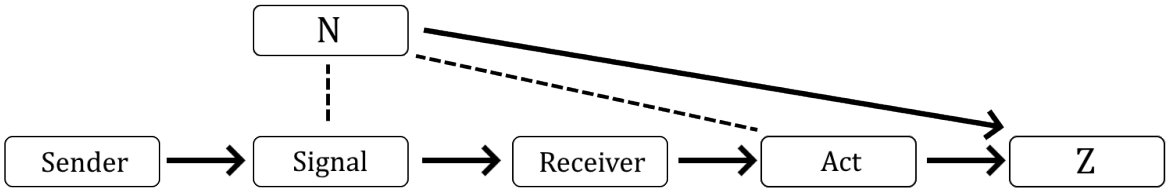

Figure 5.2: The structure that signals in a given system must have is in principle a free choice. The only demand is that the signals bear an isomorphism (dashed line between Signal and *N*) to world states *N* that preserves a certain structure of those states. The relevant structure of the world states is determined by what explains how the receiver performs *its* relational proper function, that is, how it produces an act that bears the appropriate relation to the world state (dashed line between Act and *N*). Since any system of signals with the same cardinality as world states will support many structure-preserving isomorphisms, it seems to be a very undemanding condition that such an isomorphism exist. However, signals are further constrained by what the receiver is capable of responding to. In biological situations this will typically be a strong constraint. For example, honeybees could in principle use an arbitrary system of signs to indicate locations of food in their environment. But the waggle dance is so much more efficient, and so much easier for selection to produce. As a result, the actual mapping relations of the waggle dance are not arbitrary, they are iconic.

structure to an arbitrary signalling system, it is just that it is of the lookup-table type. Only when receivers are flexible enough to interpret arbitrary signals consistently can such signals exist.

5.3 The circularity problem

Mapping relations are defined in terms of proper function and Normal explanation. In turn, proper function is defined in terms of causal effects that were selected for. But mapping relations are supposed to help explain successful proper functioning. Receiver bees are able to forage more efficiently (when they do) because the waggle dance bears a mapping relation (when it does) to a source of nectar. Mapping relations *explain* Normal successful proper functioning, but they are *defined* as those relations which in fact held when receivers successfully performed their proper functions Normally. An explanation that cites in its explanans an entity defined in terms of its explanandum is circular. Circular explanation is prohibited. What has teleosemantics to say about this?

The obvious answer (and one I shall endorse) is that instances of past success help explain current success. But this answer has been disputed. Godfrey-Smith (1996, §6) and Shea (2007) argue that teleosemantics downgrades the role of truth in explanations of success. Proponents of the theory must either accept that the role of truth is not as substantive as it is usually taken to be, or supplement the theory with a richer account of mapping relations (as Shea (2007) does).

5.3.1 The objection

Godfrey-Smith argues that the "orthodox naturalistic view" of cognition takes truth to be a property of representations that is a "fuel for success" (Godfrey-Smith, 1996, p. 171). When an organism's beliefs are true it is generally in a better position to get what it wants than if those beliefs were false (or if it had no beliefs). There are of course many caveats to this statement. Sometimes false beliefs engender behaviour that is accidentally successful. But as a general principle, the link between truth and success is implicitly endorsed via our tendency to cite true beliefs in explanations of successful behaviour. Fred was thirsty, and went to the fridge because he believed it had beer in it. How come his thirst was quenched? Because his belief was true. Theories that describe beliefs as bearing relations to world affairs are correspondence theories, and the correspondence view takes truth to be a general-purpose fuel for success.

The task of a naturalist theory, says Godfrey-Smith, is to describe the natural property that has these fuel-for-success characteristics. He argues that teleosemantics does not meet this criterion. Signals are true when they bear the same relation borne by previously successful signals of the same kind. But this definition of truth says no more than that true signals are those that fall under a certain historical pattern. Explaining successful behaviour by appeal to true signals is then no more substantive than saying 'the system did the same thing as it did on past occasions of success.' While such an explanation may not be prohibitively circular, it does not have the force that a genuine fuel-for-success approach seems to demand.

For Godfrey-Smith, Millikan's description of her theory as a version of the correspondence theory of truth is misleading. Her theory does not say, ahead of time as it were, *which* relations will count as mapping relations that help organisms achieve success. She does not

specify a pre-existing resource – a fuel for success – that can be identified with truth.

Shea (2007) agrees with Godfrey-Smith that defining mapping relations solely in terms of past success is insufficient. Shea compares explanations of success that appeal to mapping relations with explanations of the causal powers of sleeping pills that appeal to their possessing dormitive virtue. Defining 'dormitive virtue' as 'that property which has previously caused people to fall asleep in the past', such an explanation does no more than locate the present sleeping person in a historical series. Everyone in the series shares the property of having fallen asleep after having taken the pill, but nothing more substantive is given in terms of (for example) the pill's chemical constitution. For Shea, such an explanation is not as substantive as it ought to be. Appealing to mapping relations should have more explanatory power.

Both Godfrey-Smith and Shea find more force in definitions of semantic relations that include mention of information. Godfrey-Smith (1996, p. 186) mentions that Dretske's account, which includes a role for *indication* in the definition of truth conditions of a signal, avoids the problem. Shea (2007, §5.1) explicitly adds a condition to the teleosemantic definition of mapping relations that cites what he calls correlational information. In both cases, Millikan's definition is argued to be too weak, and more restrictive conditions on what gets to count as a mapping relation are offered.

5.3.2 A reply

The problem[6] reveals a significant difference between Millikan's and Godfrey-Smith's understanding of the explanatory role of semantic content. For Godfrey-Smith, semantic content

[6]Millikan (2007) and Artiga (2014b) also offer replies to the circularity problem. I intend my reply to be consistent with those accounts.

is an explanandum. The naturalist project is tasked with providing an explanans: how did content get into the world? From this perspective, teleosemantics can be seen as offering an answer to that question. Its answer is a historical, selective story about signalling behaviour. By contrast, Millikan treats semantic content as an explanans for a more basic explanandum: reliably successful functional behaviour. The naturalist project is tasked with defining, or otherwise characterising, the natural property (or relation) that best captures our pretheoretic use of 'semantic content'. From this perspective, teleosemantics can be seen as characterising that relation. I use the term 'mapping relation' for the kind of relation picked out. Now, it so happens that teleosemantics characterises that relation via its appearance in a certain *other* kind of explanation: Normal explanation.

We should therefore begin analysing the circularity problem by keeping distinct two different kinds of explanation. One is a pretheoretic explanation employed by scientists when faced with the question how signalling contributes to successful behaviour on a given occasion.[7] The explanans makes reference to the meaning or content of signals. An observer of a honeybee colony may ask: how do scouts find food at a rate better than chance? On at least some occasions, the answer is that bees who have previously visited that same food source communicate the location to other bees. The way in which location is communicated entails that different dances have different locations as contents. So our first type of explanation has the following form:

> PRETHEORETIC-EXPLANANDUM: Reliably successful functional behaviour (in particular, success of a signalling system) on a particular occasion.
> PRETHEORETIC-EXPLANANS: Signals have semantic content. When that con-

[7] By 'pretheoretic' here I mean prior to a philosophical theory of semantic content, not necessarily prior to scientific theories pertaining to signalling behaviour.

> tent is true (i.e. when the truth condition of a signal obtains), receiver behaviour can be successful at a level greater than chance.

The question the naturalist must answer is what property of signals should be identified with their having semantic content. The way teleosemantics defines semantic content is via a specific kind of theoretical device called a Normal explanation. A Normal explanation has many things in common with explanations as usually understood by scientists and philosophers. Normal explanations have an explanandum and explanans, and the facts adduced in the explanans are supposed to show how or why the phenomenon mentioned in the explanandum comes about. But Normal explanations are special in that both explanandum and explanans are constrained to be of a certain form. The explanandum is always how the ancestors of a given device performed a specified proper function that this device now has. This explanandum is relativised to a particular proper function of a particular (usually present-day) device. But the explanandum is *not* how *this* device performs the function. It is how its ancestors did it:

> NORMAL-EXPLANANDUM: Historically usual successful performance of a particular proper function (in particular, of a signalling system).

Of course there will be many different explanans across the range of different kinds of proper function. Things get interesting when we restrict our attention to sender-receiver systems. In that case, Millikan claims (and I argued in chapter 4) that the explanans must cite a relation:

> NORMAL-EXPLANANS: Signalling systems historically performed this function in part by having signals that bore particular relations to success-relevant world affairs.

The full explanans would cite how the consumer reacted to the signal, and how this

produced success in part as a consequence of the relation borne by the signal to the world affair (as told in chapter 4 and section 5.2).

Consider two contrast cases. First, imagine a hypothetical sender-receiver system that does *not* achieve success at a rate better than chance. (Suppose bees only perform dances in random directions, and so receivers never do better than if they forage alone.) Then there is nothing to explain, or certainly nothing that could appeal to a relation between a signal and a success-relevant world affair. Second, imagine a system whose ancestors all achieved success in different ways. (Suppose each generation of bees is born with randomised dancing and dance-following strategies, but by sheer chance has improved its foraging as a result.) Then there is no single set of mapping relations – because no signal mapping *rule* – that signals bear towards world affairs. That is why Normal explanations must be univocal.

Teleosemantics defines semantic content as a signal's being-supposed-to-bear-a-relation. What relation? The one picked out by the Normal explanation of its ancestors' proper functioning. When the signal bears this relation, it is true. When it does not, it is false. This is how semantic content is to be understood naturalistically, for the purposes of applying it in PRETHEORETIC-EXPLANANS.

In a sense, the teleosemantic understanding of PRETHEORETIC-EXPLANANS simply subsumes a present case of successful behaviour under a historical pattern. But it *also* gives a causal story (assuming my account in chapter 4 and section 5.2 is adequate) because it picks out the specific mapping rule signals must bear in order to cause success Normally. In other words, PRETHEORETIC-EXPLANANS picks out the specific relation that enters into the causal explanation of success of a present device. It does this because NORMAL-

EXPLANANS picks out historical ancestors of that same relation.[8]

5.3.3 A rejoinder and further reply: The origin of mapping relations

I have laid out two types of explanation in order to show how teleosemantics conceives of their relationship. But there is a third explanandum in need of an accompanying explanans. Godfrey-Smith treats semantic content as an explanandum in its own right. A Normal explanation may appeal to relations that obtained on historical occasions, but I have so far remained silent on how those relations originate. There was a time in history when there were no bee dances, hence no mapping relations of the waggle-dance variety. Nowadays, there are many such relations. Surely this proliferation calls for explanation. Let us therefore define a new explanandum:

> ORIGIN-EXPLANANDUM: The evolutionary origin of a particular set of mapping relations.

Godfrey-Smith (as I read him) suggests that any reasonable explanans will render the explanatory role of mapping relations problematic. That is because selection must surely play some role in explaining the origin of mapping relations. Yet mapping relations are supposed to play a role in explaining successful behaviour – and 'successful behaviour' in this context just means 'behaviour that gets the system selected'. If selection of a sender-receiver system explains the origin of its mapping relations, and those same mapping relations explain how the system is selected, we have a circular explanation.

It is important to note that the explanandum here is different from the two given earlier.

[8] I understand Millikan to be making the same point when she says that explaining the operation of a device in terms of its purpose (including devices whose purposes are a consequence of historical selection) "is used as a convenient way to give a definite description of the mechanism that is causally involved" (Millikan, 2007, p. 441).

PRETHEORETIC-EXPLANANDUM said nothing about semantic content. Even NORMAL-EXPLANANDUM said nothing about mapping relations. It is only when we ask about the origin of mapping relations that the threat of circularity seems to arise. Normal explanations themselves say nothing about the origins of mapping relations, just that there must actually be some relations involved if sender-receiver systems are to achieve success at a rate better than chance. If there is a circularity problem, therefore, it depends on the explanans we give for ORIGIN-EXPLANANDUM:

> ORIGIN-EXPLANANS: Mapping relations arise through selection on a sender-receiver system.

As we usually conceive of it, selection acts on things that *already exist*, preserving or destroying them. If mapping relations are selected for, they must pre-exist selection. But Godfrey-Smith (1996, p. 185) argues Millikan cannot assent to this. Millikan places no restrictions on what *can* count as a mapping relation. Any relation between a signal and the world could count, as long as it engenders success. So the teleosemantic definition of semantic content as mapping relations does not pick out some pre-existing relation between a proto-signal and the world, that then *becomes* the semantic content as a consequence of selection. Rather it identifies the relation after the fact. It is "*post hoc*, as far as explanations of success are concerned" (Godfrey-Smith, 1996, p. 188, emphasis original).

Does this definition render explanations of success circular? It does not render the pretheoretic explanation circular, because that explanandum was a *present* instance of success, and the explanans that mentions semantic content is interpreted in terms of *past* success. But doesn't this render any putative explanation of past success circular? No: for any given instance of success, there is a lineage of systems that existed prior to it, and that lineage is

what the Normal explanation (hence the definition of content) is concerned with. If we go far enough back in time, the proto-signal will barely exhibit any kind of relation that can be mentioned in a Normal explanation. But then the explanation of the system's success will not mention relations at all. That is as it should be: when we go far enough back in time, the systems that eventually become signalling systems are not yet signalling systems. Mapping relations are only available as part of a selective explanation once they have emerged from earlier forms like cue-reading and purely causal mechanisms (section 1.5). The same is true of any mechanical or relational feature of an adapted system.

In sum, separating out different possible explananda about signalling demonstrates that there is no circularity in the teleosemantic definition of mapping relations. Before concluding this section I will make the argument vivid with an analogy.

5.3.4 An analogy: the key shaped by doors

Imagine a long, long corridor with a sequence of millions of doors, each with an identical lock. A person stands at the entrance to the corridor and begins walking. When they come to the first door they take out of their pocket a thin blunt instrument made of a kind of soft metal. They use it to jimmy open the door and walk through. In so doing, the instrument gains a very faint impression of the lock's inner mechanism. At the second door, the person does the same thing, and again, by the time the door springs open, the lock has left a slightly firmer imprint on the metal. Fast forward several million doors, and the combined effect of myriad locks has worn the instrument down into what looks and works like a key. What is more, since all the doors have the same lock, each subsequent door can now be opened much more easily.

If we wanted to explain, *now*, at door five million, how the person successfully gets through the door, we would (I assume) be justified in saying they possess a key to the door. We could subsequently pose a different explanandum, namely how the person came to possess the key. The explanans would perhaps cite the original blunt instrument, plus the impact of millions and millions of doors on its form. There is a mutual interdependence between these two explanations. This interdependence is a consequence of the fact that the key is shaped by 'feedback' from each door. Each door opened made the next a little easier to open, shaping the key a little more finely into the required shape. I take it there is no reason to treat either of these explanations as circular. If we wished, we could install a micro-camera inside each lock, showing exactly how the instrument is imprinted on by the lock at each door. We would obtain a very detailed story of the origin of the key. This would not prevent us explaining how the person opens the next door by citing the key – in particular, by citing the fact that the key is a key for *this door*.

To be explicit about the analogy, the person corresponds to a lineage of sender-receiver systems, the key is a signal, the lock is a world affair, and the key/lock relationship is the signal/signified relationship.[9] Each door is a generation, and each door-opening is a collection of selective events that led to systems in the next generation being slightly better at door-opening than previously. In other words, each door-opening strengthens the key/lock relationship, just as selection can lead to more precise mapping relations.

Let us ask the question about circularity again. Does a key's being shaped by door-openings prevent it from being used to explain door openings? Obviously not, as long as you

[9]Note that I am not treating the person as sender, the key as signal, and the lock as receiver. That might be a useful analogy to make on another occasion, but I am not making it here.

do not try to explain one door-opening by reference to the key's shape *after* that door was opened. As long as temporal order is maintained, there is no circularity. That is why I have introduced and emphasised three distinct kinds of explanation. The pretheoretic explanandum is about a *present* case of signalling behaviour, corresponding to (say) door number five million. The instances of proper functional success that form part of the Normal explanandum correspond to a certain subset of previous door-openings, those for which the instrument was genuinely key-like and not just a blunt instrument. The origin explanandum asks how the instrument became key-like in the first place. The relationships between these explanations do not prevent us from citing the key-lock relation when explaining how the five millionth door is opened.

Finally, we might reasonably ask at what point the blunt instrument becomes a key. This corresponds to questions about where pre-signalling behaviours such as cue-reading end and signalling begins. I assume that in each case there is no firm threshold. After all, the same is true of most functional behaviour in biology. Natural selection is a continuous process, but it gives rise to distinct categories. We can precisely specify the mapping rules of signals like the waggle dance despite the fact that it is continuous with earlier forms of behaviour that could only reasonably be described as cue-reading.

5.4 Mapping and content

Teleosemantics identifies mapping relations with semantic content. This is among the more controversial results of the theory. One motivation for this move is that folk psychological explanations that appeal to content have a similar structure to natural science explanations

that appeal to information. Folk psychology explains behaviour by referring to the contents of mental states. Observing Fred moving towards the fridge, we might hypothesise that he both desires a beer and believes that there is one in the fridge. The fact that this content is truth-evaluable allows us to explain not just behaviour, but also success and failure. If Fred returns from the fridge empty-handed, we may posit a false belief as the reason. In traditional philosophy of mind, the canonical folk-psychological explanandum is reliably successful behaviour. Teleosemantics claims that the explanans consists of inner states that bear mapping relations, which are derived from proper functions. This explanatory structure applies in biological contexts as well. Biological entities exhibit reliably successful behaviour, and they often do this because they possess inner states that bear mapping relations to relevant distal affairs.

The view that teleosemantics applies up and down the spectrum of biological-cognitive sophistication is consistent with the idea that human psychological states are special in various ways. Sometimes teleosemantics is perceived as asserting *no explanatorily salient differences* between, say, human intentions and the waggle dance.[10] There are two reasons to reject this characterisation. First, so far as I have described the theory, teleosemantics asserts that mapping relations are a feature that biological and psychological states share. It does not preclude other features – whether properties or relations – from distinguishing them. For example, one proposed aspect of psychological states that is often cited as a desideratum of theories of mental content is *intensionality* (Hutto and Myin, 2013, p.79ff). Nothing that has

[10] I have met at least one person who took Millikan's view to be that individual beliefs are selected for genetically, and therefore the semantic properties of beliefs are exactly equivalent to the semantic properties of the waggle dance.

been said so far precludes attribution of intensionality to mental states.[11] Second, Millikan has taken great pains to distinguish human psychological (and linguistic) phenomena from biological (and less sophisticated cognitive) states.[12]

I might have dedicated a whole chapter to the differences between biological signals and philosophically paradigmatic contentful states. In my opinion the most important difference, for the purposes of explaining content-using systems, lies with **decoupled representation**. The term has two connotations. First, sophisticated signals come to be purely descriptive or directive, rather than equally both (Millikan, 2004b, Part IV). Second, sophisticated cognition enables representation of objects, events and individuals without necessarily triggering an immediate response (Sterelny, 2003, §3.1). Both decoupling directive from descriptive aspects, and decoupling recognition from response, are significant stages in the evolution of communicative and cognitive sophistication. Explanations that appeal to decoupled representations in either sense are typically aimed at *flexible success*. The idea is that the one-in-one-out profile of simple signalling systems – including those with as sophisticated 'codes' as the waggle dance – does not enable a great deal of explanatory traction. Explanations that appeal to content gain little or nothing over and above mechanical explanations of the same phenomena (Shea, 2018, p.200ff.).

There is something right about the idea that content-involving explanations gain force when systems use decoupled representations (in either of the two senses). But I maintain that the basic explanatory pattern, of appealing to mapping relations to explain success, holds

[11] For independent reasons, Millikan (2018a) rejects the claim that mental states possess intensionality.

[12] For a small sample, see the distinction drawn between intentional icons and representations in Millikan (1984, p. 96), the difference between human and animal mental representations in Millikan (2004a), and considerations of the evolution of uniquely human cognitive abilities in (Millikan, 2004b, Part IV).

even for primitive signals. It is because the resources used to transmit the signal are distinct from the resources the receiver uses to perform the relevant act that signals are apt for informational – content-involving – explanation.

Another feature of philosophically paradigmatic content is its supposed precision. The word 'dog' picks out a specific class of mammalian quadrupeds, whereas a vervet monkey's leopard alarm call may not distinguish leopards from cheetahs, or indeed any large ground predator. This observation has led to objections raised against teleosemantics such as the disjunction problem (Fodor, 1990) (Millikan, 1993, p. 7), and issues with modelling such as the partition problem (Birch, 2014b, §6). I have elected to set these issues aside, for a few reasons. First, the disjunction problem has received careful attention from Neander (2017, §§7-9), and several of her responses might be borrowable to support sender-receiver teleosemantics. Second, Griffiths (2018) argues that the disjunction problem is a conflation of three 'problems', only one of which is serious – yet even then, the serious problem is not fatal to teleosemantics. Third, Harms (2004) warns that we should not treat formal models as attempts to *translate* the contents of biological signals into human language (Griffiths (2018) appears to offer a similar warning). If this is right, we have no reason to hold up linguistic content as a standard of precision that the content of biological signals fails to meet. What is required from a specification of correctness conditions is no more than is required from a specification of proper function – and this is just the kind of assignment evolutionary biologists routinely make when they attribute adaptations. Fourth, human language and conceptual content is not particularly precise anyway: sometimes 'dog' includes wolves, sometimes it includes toy dogs. The goal of a naturalistic theory of content ought not be to demonstrate that contentful states are as precise as philosophers assume they are, but to tell

a plausible story about the evolution of contentful states that are only as precise as human thought and language in fact are. This entails figuring out just *how* precise human content is (as well as figuring out what 'precise' means in this context), but my point is that this is something we need to discover rather than assume.

5.5 Mapping relations and communication theory

5.5.1 Variant and invariant aspects

According to teleosemantics, senders are supposed to produce signals that map. A mapping is a relation that holds between a signal and its correctness conditions. Directive signals map onto states of affairs causally downstream from the receiver, that it is a proper function of the receiver to bring about. Descriptive signals map onto states of affairs that must occur in order for the receiver to perform its proper function Normally. These may or may not be causally upstream of the signal itself (figure 1.4, page 18).

Millikan (1984, 2004b) characterises mapping relations in terms of **variant and invariant aspects** of signals. Invariant aspects are the same for all signals in a system, while variant aspects differ between signals.[13] Imagine a simple alarm call that has a certain pitch and duration. These serve as its invariant aspects. Its variant aspects are the different times and locations that it is emitted. We might say that, with respect to a particular system, invariant aspects distinguish signals from non-signals while variant aspects distinguish signals from each other.

[13]This sense of 'invariant' is different from both my own and Woodward's notions of invariance discussed in section 4.2.

It may seem unintuitive that place and time should count as variant aspects. They are not intrinsic properties, certainly. Nevertheless they are physical in the sense of being specifiable in terms of physical properties and relations. For signalling systems that make use of only one type of signal, there needs to be some variant aspect for that signal to gain traction on the world. Place and time are the obvious candidates for variant aspects of simple signals. More sophisticated signals may contain further variant aspects. Consider an alarm call whose pitch varies with the proximity of a predator. Now, pitch is a variant aspect, along with the place and time of the call. The alarm call of the black-capped chickadee varies in intensity with the size of predators, and this variation is reflected in receiver behaviour (Templeton et al., 2005). Perhaps the most famous example is the honey bee waggle dance, with its spatial components indicating the direction and distance of food or viable nest sites (see the case study on page 64).

Defined this way, variant and invariant aspects are necessary for informational measurements to be made. If no variant aspect could be specified, there could be no transmission rate in a signalling system. Here I am identifying variant aspects with the possible values of X, the transmitted signal (see appendix A). If the signal had no variant aspects, there would be at most one value of X. This sole value x would have a probability of 1. The entropy of X would be $H(X) = p(x)\log\frac{1}{p(x)} = 1 \times \log 1 = 1 \times 0 = 0$. Transmission rate would therefore be $R = H(X) - H(X|Y) = 0 - 0 = 0$. This, I submit, is how to understand Millikan's variant aspects formally (at least for simple signals).

The transmission rate enabled by a code depends on (among other things) how many variant aspects the code is able to induce in its codewords. Supposing the time and place of transmission are fixed, a code with only one codeword could not support any transmission

at all. A code with two codewords enables a rate of at most one bit per codeword; and so on. Paul Revere's apocryphal signal differentiated between two states of the world: that the British were approaching by land or by sea. It carried one bit of information.

5.5.2 Mapping relations are neither codes nor strategies

In section 3.2 we saw that codes are a special case of strategies. How do they relate to mapping relations?

To understand how mappings differ from strategies and codes, recall how each has been defined. A strategy is a piece of selected behaviour, typically called an adaptation in biology, that is formalised in game theory. A code is a special case of a pair of strategies that is formalised in communication theory. Mapping relations are relations that may or may not hold between signals and world states and are cited in causal explanations of Normally successful behaviour of receivers.

In the central model, encodings ensure mapping relations are borne between source strings and encoded strings. The causal *Source string-Encoded string* relation runs in tandem with the descriptive mapping relation, which explains how the receiver (decoder) achieves success (reconstructs the string) at a rate better than chance. Conversely, the decoding relation runs in tandem with the directive mapping relation. This is a special case. Rarely does a signal map exclusively to one thing that immediately causally precedes it and one thing that is immediately causally subsequent. Figures 5.3 and 5.4 depict this situation for the basic sender-receiver model and the central model.

It is the mapping relation between source string and encoded string that explains the receiver's success in the first instance, not their causal relation. We can see this by exposing a

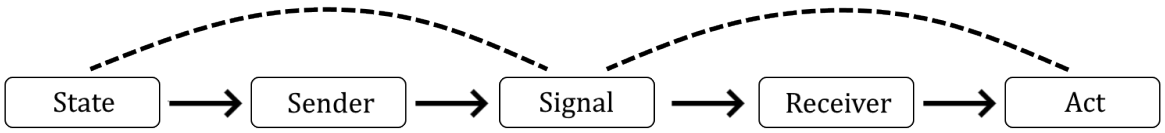

Figure 5.3: Mapping relations for a basic sender-receiver model. The only goal of the system is to match acts to states. Compare figure 1.4, page 18, which depicts the general case in which the descriptive correctness condition is not causally upstream of the signal, and figure 5.4 which depicts mapping relations in the central model.

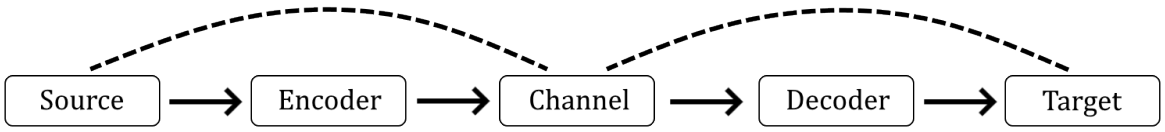

Figure 5.4: Mapping relations in the central model. The goal of the system is to match target string to source string. Compare figures 1.4 (page 18) and 5.3.

pattern of invariance between source string and encoded string. Intervening on both strings has two interesting consequences. First, the causal connection between the two strings is broken. That is because, in general, intervening on a variable breaks upstream causal connections. Intervening on the encoded string therefore breaks its causal connection with the source string. Second, patterns of invariance in the source string-encoded string relation can be seen to explain receiver success. Recall, the receiver's job is just to produce a string identical to the source string. If there is no longer any causal connection, yet the receiver is still successful, it must be the pattern of invariance (hence the mapping relation) that is explaining success.

5.5.3 Compression and transmission

> There is a duality between the problems of data compression and data transmission. During compression, we remove all the redundancy in the data to form the most compressed version possible, whereas during data transmission, we add redundancy in a controlled fashion to combat errors in the channel.
>
> Cover and Thomas (2006, p. 184)

> There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel. [...] This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.
>
> Shannon (1959, p. 350)

In line with our teleosemantic interpretation of communication theory, I shall argue that the duality between compression and transmission can be understood in terms of the duality between descriptive and directive mapping relations.

So far we have portrayed encoding as a strategy that converts source strings into signals. A more sophisticated encoding process, that takes advantage of all the ingenuity of engineers, would separate this strategy into two. First, it would throw out redundant aspects of the original string that are not necessary to reconstruction. Communication takes effort, and it would be better value for money to transmit as short a signal as possible. Second, the encoding process should add some kind of redundancy to the message so that a certain amount of interference will not threaten attempts to reconstruct it.

The part of the encoder that does the first job is called the *source coder* (it comes right after the source). Source coders remove redundancy from symbol strings. Suppose your string is five hundred *A*s followed by five hundred *B*s. It would be a waste of space to

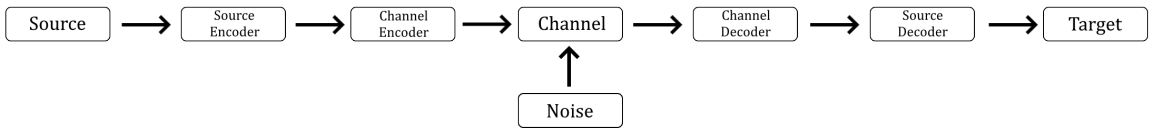

Figure 5.5: The extended central model. The source encoder removes redundancies in source strings, while the channel encoder judiciously *adds* redundancy to counter noise in the channel.

transmit all one thousand symbols if there was a shorter way of transmitting the description *five hundred As followed by five hundred Bs*. This first stage of encoding responds to the pressure to save on signalling resources.

The second part of the encoder, called the *channel coder* (it comes right before the channel), reflects a competing pressure. No matter how pristine your telegraph wire, cosmic rays and other inconveniences will interfere with electronic pulses sent along it. Strings that live in Platonic heaven are free from interference, but lowly human-made signals cannot avoid noise. To counteract noise you must judiciously insert redundant aspects into your message. Suppose you send the same signal five times. Interference might affect each of those five copies in different ways, but by taking the 'average' over all five received signals, the decoder might well be able to recover the original string. Sending *five hundred As followed by five hundred Bs* five times might take less time and power then sending the entire sequence once. It will certainly be more efficient than sending the entire sequence five times. The central model can be extended by distinguishing source and channel coding (figure 5.5).

Less brute-force methods of ensuring redundancy have been developed and are staples of introductory courses in communication theory. Suppose you want to send a four-symbol string in which each symbol can be A or B. You can make your channel coder add three

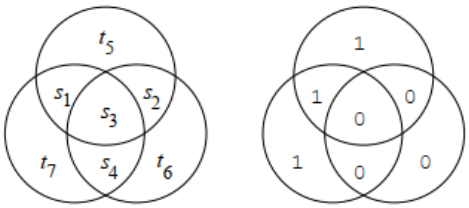

Figure 5.6: How the Hamming (7,4) code works. Taken from MacKay (2003, p. 9), the original caption reads: "We arrange the seven transmitted bits in three intersecting circles. The first four transmitted bits, $t_1t_2t_3t_4$, are set equal to the four source bits, $s_1s_2s_3s_4$. The parity-check bits $t_5t_6t_7$ are set so that the parity within each circle is even: the first parity-check bit is the parity of the first three source bits (that is, it is 0 if the sum of those bits is even, and 1 if the sum is odd); the second is the parity of the last three; and the third parity bit is the parity of source bits one, three and four."

symbols to this string such that it is always reconstructed correctly, even if one of the resultant seven symbols is flipped by noise in the channel. The first check-symbol, which will become the fifth symbol of the transmitted string, is set to *A* iff an odd number of the first, second and third symbols is *A*. That means the decoder knows to expect an even number of As from the first, second, third and fifth symbols of the string. Then do the same thing with the second and third check-symbols, with different combinations of the original symbols (see figure 5.6). The result is a pattern of symbols whose original state can be recovered correctly even if one of the resultant seven symbols is altered during transmission. (As an added bonus, if *two* symbols are changed in transmission, the decoder knows that something has gone wrong though they won't know what it is. If *three* symbols are changed, however, the decoder cannot know that anything has changed.)

Interestingly and not always noticed is that several common forms of channel coding rely on mapping relations between symbols *within* the encoded message. The channel coder

ensures relations hold between symbols of the code string so as to enable accurate reconstruction of the message. The code just described, which is called a Hamming (7,4) code after its inventor Richard Hamming, is a basic, yet instructive, example. The relations induced between original symbols and check-symbols are mapping relations. They tell the decoder what combination of symbols is supposed to exist. The decoder can inspect the combination and confirm whether or not that combination does in fact exist, fixing mistakes when they arise. In other words, if the symbols do not map as they are supposed to, the receiver can identify and rectify the mistake. In Millikan's terminology, a Hamming (7,4) code string possesses many self-signs, symbols that represent themselves as well as other things (Millikan, 2017, §9). They are micro-signs, fine-grained and localised and doing a very specific job.

By gluing together the source coder and channel coder we get a coding scheme. Correspondingly, the receiver comprises a channel decoder followed by a source decoder. Now we can understand Shannon's cryptic remark about knowing the past and controlling the future. Source coding describes the optimal way to represent ('know') a string. Channel coding describes the optimal way to reproduce a string at the receiver – to 'control' the downstream operation.

5.5.4 Division of labour in animal signalling

In light of the division of labour between compression and transmission, consider animal signalling. The states of the world about which animals have to signal are often very simple indeed. They do not require finely articulated signals. As well as familiar subject matters of nearby danger or readiness to mate, there are territorial claims and pre-fight preparations. Signals that pick out these states need not be compressed, because there are so few possibili-

ties in the first place. The details of such events are usually not complex enough to warrant compression. Therefore, by associating encoding with sender strategies, most of the signal design corresponds to channel coding rather than source coding. Most animal signals will be geared towards adding redundancy for transmission rather than removing it for compression.[14]

Most animal signals are sent between animals with imperfectly aligned interests. Often, overt signalling behaviour will be geared towards producing effects that benefit the sender but not the receiver. Suppose overt behaviour can evolve faster than perceptual discrimination. Then manipulative senders will tend to stay a step ahead of discerning receivers. If this is right, the sender's problem will be equivalent to that of the channel coder (as opposed to the source coder). The sender has a response they want to produce, and their problem is not compression but transmission: ensuring the response gets performed no matter what. By the same token, the signal is more directive than descriptive. Those behavioural ecologists who follow Dawkins, Krebs, Owren, Rendall and Ryan in claiming most animal signals skew towards the influential end of the spectrum may well be right. This does not entail those signals are out of reach of informational formalism. That a signal is designed to control rather than represent does not place it out of reach of mathematics (see also section 7.5).

5.6 Projection and the containment metaphor

When we speak casually, we often treat relations as if they were properties. In everyday speech this rarely causes problems. But those ersatz properties are sometimes made to do

[14]The source/channel coding distinction is relevantly similar to the strategic/efficacy design distinction; see sections 1.4 and 7.2.

explanatory work. Then the question arises, are such properties appropriate for scientific theorising?

Philosophers have complained about inapt metaphors applied to signals. The relation between signal and signified is often murky, leading to ambiguous explanations. These objections centre on a practice Godfrey-Smith calls the metaphor of containment:

> The familiar philosophers' language of 'content' is potentially misleading, as it suggests that a meaning is either in a sign or not in it. The metaphor of containment is probably a bad one. A signal may have many relations to the world, relevant in different contexts.
>
> Godfrey-Smith (2011, p. 1292)

The containment metaphor is operative when a signified is treated as if it were literally contained within the signal. Describing the signified as the *content* of a signal, where 'content' is read as if describing objects in a box or liquid in a bucket, is an example. Signifieds are neither contained within signals nor properties of them. Yet scientists and philosophers routinely speak and write as if one or the other were the case.[15]

There are special cases in which signals or parts of signals are physically contained within other things. Suppose I hand my partner a glass of orange juice, intending for them to stop drinking wine. The liquid contained within the glass can constitute a signal that they shouldn't drink any more alcohol. But the glass-liquid relationship is not the same relation as that between the liquid and the state of affairs that they shouldn't drink more alcohol. A

[15] An examiner suggested a simpler response to the complaint that content attribution is metaphorical, as an alternative to the response I offer in this section. The idea would be that content and related concepts are indispensable for cognitive science, and that biology (and presumably the functional sciences in general) gets along very well by redefining what were formerly metaphors as technical terms. For example, the terms cell, adaptation, cooperation, altruism, and function were all metaphorical when originally applied to biological contexts. Over time, their usage became literal, as their definitions became fixed and agreed upon. I don't deny that this strategy would work. But I think my assertions in this section offer insights worth having, too.

world affair that is physically distinct from a signal cannot literally be a property of that signal, yet our typical patterns of speech make it seem as if they are. Can we make sense of such linguistic practices?

Several philosophers invoke the containment metaphor to explain (and, typically, criticise) these patterns of speech. Rendall et al. (2009, Appendix) object to these linguistic patterns in explanations of animal communication. Though they use the term *conduit* metaphor, the explanatory pattern they are objecting to is equivalent to the containment metaphor. They object that "this metaphorical approach also carries with it a definitional sleight-of-hand that undermines its scientific integrity" (Rendall et al., 2009, p. 240). By failing to adequately define the substance supposedly transmitted along the conduit – contained within the signal –, researchers leave a significant explanatory gap. The authors conclude that explanations making use of this metaphor should be discarded.

Rendall and colleagues build on a famous demonstration by Reddy (1979) that English speakers commonly draw on the conduit metaphor when describing linguistic meaning. Reddy's appendix includes hundreds of examples of everyday constructions that rely on the conduit metaphor. Strangely, Reddy's depiction of the metaphor *as* a metaphor leads him to endorse another kind of view of language that was soon to be criticised: the code model. Reddy (1979, p. 287) says that language "seems rather to help one person to construct out of his own stock of mental stuff something like a replica, or copy, of someone else's thoughts - a replica which can be more or less accurate, depending on many factors." This view, under the label of the *code model* of communication, was later rejected by Sperber and Wilson (1986). Indeed, Rendall and colleagues' reasons for rejecting the conduit metaphor are equally good reasons to reject the code model. Opinion among contemporary linguists is ecumenical: dif-

ferent models serve different purposes. None are exclusively correct, but none are forbidden.

Along similar lines, Hutto and Myin (2013) argue that the containment metaphor leads cognitive scientists astray. In explaining cognition, scientists often advert to information being "extracted, retrieved, picked up, fused, bounded up, integrated, brought together, stored, used for later processing, and so on and so forth" (Hutto and Myin, 2013, p. 63). Hutto & Myin reject that such claims can be literally true and, like Rendall and colleagues in the context of animal signalling, claim they are at best metaphorical. Their suggested revision to explanatory practice allows that only physical objects are manipulated. Scientists have confused the transmission and transformation of physical objects for the containment and conveyance of some abstract substance.

It seems to me that these authors are mostly right in their analysis of explanatory patterns in biology and cognitive science. But calling the practice metaphorical and leaving it at that is too quick. One reason for caution is that philosophers are apt to associate metaphor with fiction. Levy (2011), for example, uses the terms interchangeably (though he draws on a reasonably sophisticated account of fiction due to Walton (1990)). We need to be a bit more subtle. Mapping relations are not properties of signals – signifieds are not 'contained within' signals – but there is a legitimate way of describing a relation as a property, a way I will call **projection**. To denounce the practice of projection as metaphorical is to ignore other linguistic strategies, and to miss the literal truth they rely upon. The problem is not just that scientists confuse the vehicle/content distinction, or illicitly borrow a model from communications engineering. It is that we do not know how to think about relations, so we often choose to treat them as properties instead.

When we use projection, we take what is strictly a relation and describe it as a property

of one of the relata. The linguistic phenomenon by which relations come to be described as properties or contents is not metaphor – at least, not metaphor in the pejorative sense that philosophers of science tend to invoke. Treating a relation as a property of one of the relata is a verbal or conceptual shift that is more or less apt depending on how 'fixed' the other relatum is. This is more like synecdoche or metonymy than metaphor.[16] For example, I am a brother, my sister is a mother, her children are both nephews. We can define 'brotherhood', 'motherhood' and 'nephewhood' if we like, abstracting away from the diverse relata that make people brothers and mothers and nephews. When it does not matter to whom our focal agent bears the relevant relation, the agent can be treated as possessing a property instead. In other words, when relata can reasonably be treated as invariant, relations can reasonably be treated like properties. In this case we treat a relatum as representative of the whole relation. It is acceptable so long as context makes clear either what the other relata are, or that their identities are irrelevant to the point being made.

Hofstadter (1979, pp. 171–2) offers an analogy between meaning and weight. Your bathroom scales do not measure mass, despite showing quantities in kilograms. They measure weight. On the surface of the Earth, weight is converted from Newtons of force into kilograms of mass thanks to gravitational acceleration serving as a constant conversion factor. Weight is a relation between two bodies, but when one of those bodies is assumed to be fixed (the Earth's gravitational pull is a fixed feature of most bathroom scales' environments) we can bracket the second relatum and think of the relation as a property. Dretske (1981, p. 80)

[16] Synechdoche: "A figure of speech in which a more inclusive term is used for a less inclusive one or vice versa, as a whole for a part or a part for a whole"; metonymy: "(A figure of speech characterized by) the action of substituting for a word or phrase denoting an object, action, institution, etc., a word or phrase denoting a property or something associated with it, e.g. as when referring to the monarchy as 'the crown' or the theatre as 'the stage'; an instance of this" (both from Oxford English Dictionary, 2020).

and Haugeland (1998, p. 360 n. 28) make much the same point.

In general, fixing one or more relata allows a relation to be viewed as a property. This is projection. When conceived as a property of an object, weight is a projection. 'Being a sibling' is a projection from the relation between siblings to a property of an individual. To take a more scientific example, sufficiency (as in the phrase 'sufficient statistic') is a property of a statistic with respect to the sample from which it was calculated. A statistic is sufficient when no further information can be calculated from the sample that cannot also be derived from the statistic itself. Sufficiency, as a property of a statistic with respect to a sample, is therefore a projection.

Analysing properties as relations in disguise has philosophical pedigree. Relativist construals of probability and value treat these not as intrinsic properties of events or objects, but as properties borne only with respect to an agent's perspective. The fitting attitude account of probability and value is an example (Rabinowicz, 2015). However, the majority of such accounts relativise properties to agents' mental states: they are *subjectivist* theories. In the present project, I have argued for a relativist reading of properties ascribed to signals, but with respect to *proper functions* rather than mental states. I suspect that many subjectivist accounts of philosophically interesting properties have a relativist version. Subjectivist theories' relativist counterparts result from discarding the agent's mental states and retaining only its proper functions.

Projection is not metaphor. It is a pattern of speech. The relations and ersatz properties involved are more closely linked than the source and target of a typical metaphor. It would be wrong to assume that any attribution of properties is either literal or figurative, as though there were no other kind of linguistic strategy. Furthermore, words change their meanings

based on usage. If we were to consider as literal only the original meanings of English language terms, every utterance of every scientist would be metaphorical: scientists cannot literally refer to *electrons*, because the word 'electron' derives from a Latin word meaning amber-like (and a Greek suffix) and the elementary particle to which the word refers is nothing like amber. If that argument is incoherent, then so is the claim that scientists cannot literally appeal to semantic content. If there are such things as mapping relations, then they are what is being literally referred to by semantic content talk. This is true even if the surface grammar of content-talk attributes a property rather than a relation.[17]

Up to now we have considered cases in which the relation between a signal and its signified has been grammatically marked as subject-predicate. For example, 'the content of the signal is *X*' takes the content of the signal as subject and predicates *X* of it. There is also the case where the signal/signified relation is marked as subject-verb-object. In describing and explaining the cognitive state of a frog as it reacts to the appearance of a fly in its immediate environment, we might say *the frog's neural state signifies the fly*. This should be (and presumably is) regarded as equally dubious by opponents of the containment metaphor. A signal cannot really *signify* its signified, if signification is an action. Signification (representation, registration, indication) is not an act performed upon an object by a signal.

One benefit of the subject-verb-object construction is that it highlights the asymmetry in the sign-signified relation. The appropriate analysis is probably the same as that argued above. What is being stated is the existence of an asymmetric relation between signal and

[17] What happens when we legitimately use projection and then capture the resulting property in a causal model? Wouldn't that lead to a causal role for content, contra my claims in section 4.2? Yes, and this is all the more reason *not* to use projections in causal models. This seems to be what has happened, for example, in the account sketched by O'Brien and Opie (2015) (and perhaps also Cummins (1996)). For them, content explains proximate behaviour primarily and success secondarily; for teleosemantics it is the other way round.

signified. In English, our two main tactics for expressing such a relation are the projection tactic (leading to a subject-predicate construction) and what might be called the action tactic (leading to a subject-verb-object construction). Short of extending the language, it is not clear what more the sceptic can reasonably ask of the realist.

In a system with network structure, the containment pattern gives rise to a processing pattern. In a signal processing system like a computer, different electrical impulses are produced, transmitted, destroyed or transformed. Hutto and Myin (2013, p. 63) point out that it is only the signals, not the signifieds, that are having these things done to them. But they do not tell us what is happening to the relevant mapping relations. Why not say a relation is transformed or destroyed when one of its relata (supposing there to be only two) is transformed or destroyed? To reject this would require one to adopt a strange linguistic prescription. You must demand that every verb in your theory apply literally to objects and properties and only metaphorically to relations. I do not believe this is a good rule to live by. I do not think we should exclude relations from our explanatory repertoire. If properties explain, so do relations. Sometimes, making sense of scientific explanation requires taking note of the grammatical constraints scientists face. Until we have a formal account of explanation we must remain, as the saying goes, eternally vigilant.

5.7 Conclusion

Mapping relations promise to illuminate explanatory patterns shared by signals in the central model and biological signals. Problems interpreting teleosemantic theory can be resolved by paying close attention to different explananda. The teleosemantic distinction between

directive and descriptive mapping relations corresponds to the duality between compression and transmission. And the question how semantic content can be a 'property' of signals should be answered by paying attention to the different linguistic strategies available for expressing relations between signal and signified.

Chapter 6

Sense and sense stability

6.1 Introduction

The problem addressed in this chapter is how teleosemantics can accommodate the content of signals sent between uncooperative agents. I argue that in light of recent work, prospects are good for formalising teleosemantics in a manner that adequately captures conflictual content.

According to teleosemantics, states and processes bear meaning when they stand as intermediaries between cooperating senders and receivers. Cooperation results from coselection or joint design. Signal content is defined as the mapping relation(s) that explain system success. Teleosemantics appears to assert a strong link between correctness and cooperation.

However, signalling in nature is rarely (if ever) a consequence of perfect coadaptation. Except perhaps for certain cases of intra-organismal signalling characterised by extreme common interest, selection provides incentives for signallers to exploit or deceive each other. Teleosemantics describes an ideal which nature fails to live up to. Proponents of the theory hope that signals in uncooperative contexts can be attributed meaning based on (possibly hypothetical) perfectly cooperative contexts. For example, Millikan (2004b) considers systems with the function of producing false representations. She argues that the contents of these representations must derive from a prior cooperative mechanism:

> [M]any biological systems ride piggyback on systems developed earlier for

> other purposes. Systems whose jobs were to distort certain beliefs would have to ride on more general systems whose basic jobs were to produce true beliefs. Otherwise there would be no standard mapping rules according to which the distorted beliefs were designed to map world affairs hence according to which they were false.
>
> Millikan (2004b, p. 86)

The thought is that by defining content with respect to cooperative systems, we can assign meanings to uncooperative signals based on their cooperative counterparts. Intuitively, the bad cases are derivative on the good.

Success on this view is defined in terms of *stability*: signals serve joint proper functions and a joint proper function is a stable interaction between two or more agents. Correctness conditions are just those states of affairs that promote the interaction recurring. In short, it is more accurate to say that teleosemantics asserts a strong link between correctness and *stability*. However, it is not easy to cash out this intuition in detail.

Fortunately, two distinct mathematical approaches also link correctness and stability, and can be used to describe systems that lack full cooperation. In this chapter I argue that **costly signalling theory**, understood in its most general form, provides resources to support a teleosemantic interpretation of signalling with divergent interests. That is because costly signalling theory applies more broadly than the economic and biological contexts in which it was first proposed. Signalling is costly no matter who does it. It is costly because all behaviour is costly, and signalling is a kind of behaviour. Formulating and sending signals takes energy and time. Receiving and responding to signals takes energy and time too. Physiological and opportunity costs are borne by all senders and all receivers. No matter who, how or why, signalling is costly.

But there are benefits that offset these costs, and a fundamental way to explain signalling behaviour points toward the trade-off. All signalling systems that bear costs can be analysed within a common framework. The branch of engineering called **rate-distortion theory** employs the same framework. It delivers results about optimal signal strength and transmission rate with respect to communicative goals. Happily, costly signalling theory has produced similar results, though their kinship seems to be neglected. To my knowledge, only de Polavieja (2002) and Martínez (2019) explicitly mention the links between them. Yet, I will argue, their relationship dramatically improves the prospects of a teleosemantic approach to meaning.

This chapter demonstrates how to expand teleosemantics to encompass signalling partnerships with divergent interests. The chapter proceeds as follows. I raise the problem (section 6.2) and survey existing proposals due to Artiga (2014a) and Shea et al. (2017) (section 6.3). With the use of a generalised sender-receiver model (section 6.4), I describe the common structure of teleosemantics, costly signalling theory and rate-distortion theory (section 6.5). Then, I argue that functional content is a reasonable formalisation of teleosemantics that attributes content in systems with divergent interests (section 6.6). This could be viewed as either a novel extension to, or simply an interpretation of, teleosemantics. Toward the end of the chapter, I respond to an objection (section 6.7).

6.2 The problem of conflict

We live in a disputational world. A multitude of beings scramble for finite resources. Conflict is everywhere; common interest is rare. The energy bound up in an organism's body is at

risk of being metabolised by predators. Even creatures not actively trying to destroy each other are looking to gain an advantage by exploiting their fellows. Many kinds of behaviour serve the antagonistic ends of biological agents, including signalling. Consider the case of fireflies of the genus *Photinus*, cited by Stegmann (2009, §2), Skyrms (2010, §6) and Artiga (2014a, p. 362ff.), and described in the following case study.

Firefly predation

Many fireflies use flash patterns to signal to conspecifics, especially during mating season. Members of the genus *Photinus* display the following courtship behaviour. Males flash a species-specific pattern in an area where one might reasonably expect females to be. Upon receiving the signal, and if she is willing to mate, the female flashes back her own pattern. The character of the flash pattern is important for species recognition (Lloyd, 1975). If he receives a favourable response, the male approaches the female to mate.

All seems well with this communication system. Especially because different species emit distinct patterns, the risk of accidental crosstalk is minimal. Unfortunately, interference comes from a different genus of firefly. Females of the species *Photuris versicolor* have a vested interest in tapping the line of *Photinus* and other susceptible prey. The predator answers male flashes by emitting a reasonable imitation of the female response. When the male approaches, the predator catches and eats him.

The case of the fireflies, and many more like it in nature, poses a challenge to sender-receiver teleosemantics. The content of a signal is the state of the world picked out by its mapping rule. Mapping rules are determined by coevolution between sender and receiver.

Content is defined in terms of historical cooperation. In the firefly case, the predator is being decidedly *un*cooperative. So, teleosemantics seemingly cannot ascribe content to the predator flash.

On the face of it this conclusion might not trouble us: without cooperation there is no joint success to explain, so no work for content to do. There is no reason to posit content because there is no explanatory burden for it to bear. Nothing is lost, therefore, by taking such signals as *P. versicolor*'s imitative flash to be contentless.

We will, however, face trouble when trying to accommodate this result with scientific practice. Biologists routinely offer explanations for conflictual signalling in terms of content. They attribute the same properties to conflictual and cooperative signals. One marker for such practice is the use of terms like 'deception'. Deception usually implies false content, and this is indeed how biologists use it. A recent spurt of papers seek to determine what else the concept entails (Martínez, 2015, 2019; McWhirter, 2016), but those issues are not relevant here.

Take the firefly case. Lloyd (1975, p. 453) describes the flash pattern of the predatory firefly *P. versicolor* as a "false signal", as it mimics the flash patterns of its prey species. Many other descriptions of biological signals under conflict appeal to their putative content.

So we are left with a dilemma. We must either show that teleosemantics attributes content in these cases, or advance a revisionary account of a popular scientific practice.

At this point Millikan's own account of uncooperative senders seems promising. The predator's deceptive flashing pattern "rides piggyback" on the cooperative flash system (Millikan, 2004b, p. 86). The deceptive flash has content because it mimics a cooperative flash, whose content is determined according to the standard teleosemantic story. This is, more or

less, the account I shall endorse. However, there is a lot more detail to work through before this line of thought can be accepted. Millikan tends to mention these kinds of cases only in passing. She does not dedicate the same amount of precision to uncooperative signals as can be found in the paradigmatic teleosemantic account of cooperative signalling.

To develop teleosemantics into a theory that can account for conflictual signalling, it will be useful to show how content is determined for signals across a range of contexts of greater and lesser cooperation. I argue that recent formal work promises to help teleosemantics meet this challenge. Moreover, expanding the remit of teleosemantics to include partial cooperation, and even extreme conflict, reveals further coherence with the formalism of communication theory. The rest of this chapter argues for these claims. First, I survey recent work that aims to solve the puzzle.

6.3 Existing solutions

6.3.1 Artiga's solution

Artiga (2014a) reconstructs the problem as an inconsistent triad:

- TELEOSEMANTICS: roughly, the claim that sender-receiver teleosemantics defines representational content (Artiga, 2014a, p. 360).
- MIMICRY: "The light emitted by [Photuris] is a signal, which means something like *[female Photinus] willing to mate*" (Artiga, 2014a, p. 363).
- INCOMPATIBILITY: TELEOSEMANTICS is incompatible with MIMICRY (Artiga, 2014a, p. 363).

At least one of the three must be given up or amended. Artiga explores two responses in

depth: rejecting MIMICRY and rejecting INCOMPATIBILITY.

First consider rejecting MIMICRY. On this option, the predator's flash is not in fact contentful. Explanations of receiver behaviour that advert to relations supposed to be borne between signal and world are faulty, because there are no mapping rules underpinning them. A different explanation should be offered for the receiver's behaviour.

How can we explain why the male approaches upon seeing the flash, if not by reference to the flash's content? We could point to the physical properties of the signal. The predator flash is similar enough in form to female flashes that elicit the same response. It need not possess content to do so. A token flash is a triggering cause: we advert only to its form when explaining behaviour. And when it comes to explaining receiver failure, we advert not to false content but to the fact that the signal had no content at all.

By rejecting MIMICRY, we assert that the phenomenon of mimicry is the reproduction of form only. To take a different example, warning colouration on edible frogs is not false content. It contentlessly reproduces colouration that in another context has content and is true. This move is revisionary, although it doesn't seem drastically so. Rather, the interpretation of 'mimicry', 'deception' and so on changes. The revision is not to the usage of terms, but their interpretation.

One point in favour of this move is that it mirrors our response to the problem of non-cues (section 1.3). Recall, tokens that are not produced by the same natural processes as cues of type *C* are not cues (or at least not cues of type *C*) and so do not bear the appropriate correspondence. The same can be said here. Tokens that are not produced as part of the same signalling system are not signals (or at least not the *same* signals) and so do not have the same content.

However, there are several problems with rejecting MIMICRY. First, Artiga notes that it seems ad hoc. Revising biological explanation for philosophical purposes requires strong justification. If there is no good reason to reject mimicry except to save teleosemantics, then teleosemantics had better be an extremely well supported theory – not just philosophically, but by the lights of science as well. Perhaps such a route is viable, but it is probably best to keep it as a last resort. Second, MIMICRY embodies the fundamental principle that truth explains success and falsity explains failure. If we reject it, we forego the ability to appeal to falsity in explaining failure. And this is tantamount to admitting that we do not need to appeal to truth to explain successfully coordinated action. If male behaviour upon receipt of predator flash is explained by that signal's form only, why should we explain male behaviour in response to the *female* flash in terms of that signal's content? Contentless explanation metastasises. This is bad news for teleosemantics. In what follows, I leave aside the possibility of rejecting MIMICRY.

A more hopeful option is rejecting INCOMPATIBILITY. There are two ways this might go. First, the predator's signal system can be considered a member of the same biological kind as the female's signal system. Then there is just one sender type, and it has partial common interest with the receiver. Second, although predator and female are different senders, the signals they produce can be considered members of the same kind. Then they have the same content because they have the same function.

I will consider the ramifications of both options in a moment. Before that we will look at a more formal way of describing content in signalling systems with divergent interests.

6.3.2 Shea and colleagues' solution

While Artiga gives informal arguments supporting teleosemantics, another way to explore the same issues is through the lens of formal models. Shea and colleagues aim to expand the formal repertoire provided by Skyrms (2010). Skyrms's statistical content captures correlational relationships between signals and world affairs. Shea et al. (2017) define **functional content**, geared towards capturing conditions that contribute to stability of a signalling system. This explanatory role was earlier pointed out by Godfrey-Smith:

> A signal might do little to the probability of w_1, and raise the probability of w_2 much more, while it is the link to w_1 that explains why the relevant sender and receiver dispositions have been stabilized. Then the signal has a special kind of involvement with w_1, despite the weak probabilistic connection.
>
> Godfrey-Smith (2011, p. 1292), variables changed for consistency

The idea is to pick out conditions that were present when a signal was sent, which together with the receiver's behaviour upon receipt of that signal, contributed to stability of sender and receiver behaviour. Both statistical and functional content are vectors that describe, for a given signal, relationships between it and world affairs. Statistical content captures probabilistic relationships, saying how likely a world affair is given that a signal was sent. In contrast, functional content captures relationships of functional relevance, capturing "the degree of involvement of that state in the stabilization of the sender's and receiver's behaviours regarding that signal" (Shea et al., 2017, p. 8).

Here is the thought behind functional content (see appendix C for details on how to calculate the vector). Focusing just on descriptive correctness conditions, a behavioural response is said to be stabilised by a state of affairs when that behaviour together with that state of affairs brings a greater than baseline payoff. Baseline payoff is what signallers

would have enjoyed had the receiver acted without conditioning its behaviour on the signal. Therefore, signals that prompt beneficial responses continue to promote those responses in future. Crucially, because a certain world state or states played a role in that stabilisation, a relation can be drawn between those states and the signal.

An example helps distinguish functional content from statistical content. Consider the following game:

$$P(W) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \quad Q_s = Q_r = \begin{pmatrix} 7 & 0 & 2 \\ 4 & 6 & 0 \\ 0 & 5 & 10 \end{pmatrix}, \quad m = 2 \tag{6.1}$$

In the absence of signalling, the receiver's best act is a_3. The baseline payoff is $\left(\frac{1}{3} \times 2\right) + \left(\frac{1}{3} \times 0\right) + \left(\frac{1}{3} \times 10\right) = \frac{12}{3} = 4$. Because there are only two signals available, the players are restricted to a bottleneck like that depicted in figure 6.1. Two signals are used to distinguish three states. What is the functional content of s_1 and s_2?

First look at s_2. It is used only in w_3 and drives only a_3. The first two components of the functional content vector are therefore zero, because the signal is never the result of the corresponding states (formally, $p(w_1|s_2) = p(w_2|s_2) = 0$). The third component of the functional content vector depends on the payoff achieved in state w_3. In this case $q^s_{33} = q^r_{33} = 10$, which exceeds the baseline (of 4) by 6. The raw components of functional content are therefore $(0,0,6)$. These values are normalised to give the functional content vector $F(s_2) =< 0,0,1 >$.

Now consider s_1. The third component of its functional content vector is zero because the signal is never the result of state 3 (formally, $p(w_3|s_1) = 0$). And the second component

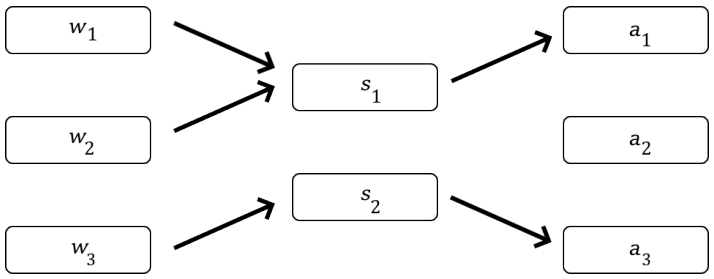

Figure 6.1: A bottleneck system in which functional content differs from correlational content (Shea et al., 2017, fig. 3 p. 14). Sender and receiver payoffs are the same. (Note that **this is not strictly a causal model**: here, nodes represent *values* of variables rather than variables themselves.)

is also zero, because the relative payoff of acts in that state does not exceed the baseline (formally, $\Sigma_k p(a_k|s_1)q_{2k} = (1\times 4+0\times 6 = 4 \leq 4)$. The first component depends on how much the payoff achieved in state w_1 exceeds the baseline, and the probability that w_1 is actual given s_1 was received: $p(w_1|s_1)(q_{11}-\overline{q}) = \frac{1}{2}\times(7-4) = \frac{1}{2}\times 3 = \frac{3}{2}$. The raw components of functional content are therefore $(\frac{3}{2},0,0)$. This vector is normalised so the functional content vector is $F(s_1) =< 1,0,0 >$.

So the functional content of the first signal is $< 1,\ 0,\ 0 >$, and of the second $< 0,\ 0,\ 1 >$. However, the first signal is also sent in the presence of state 2. Since states 1 and 2 are equiprobable, the statistical content of the first signal is $< \frac{1}{2},\ \frac{1}{2},\ 0 >$, which differs from its functional content.

Returning to the fireflies, functional content is easily calculated once the scenario is modelled. We have two states corresponding to female and predator, two signals corresponding to flash and no flash, and two acts standing for 'approach' and 'stay away'. Payoffs can be set as follows:

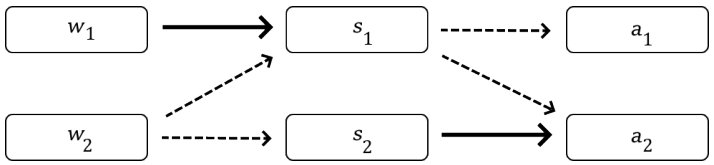

Figure 6.2: A hybrid equilibrium in the deception game (adapted from Shea et al. (2017, fig. 5 p. 17). As a representation of the firefly example, w_1 stands for a female firefly being present and w_2 stands for a predator being present. s_1 is the species-specific flash pattern, while s_2 stands for an absent signal. The male firefly can either approach (a_1) or ignore (a_2). Solid arrows indicate a deterministic strategy – females always flash, males always ignore an absent flash – while dashed arrows indicate a mixed strategy. (Again, **this is not strictly a causal model**: here, nodes represent *values* of variables rather than variables themselves.)

$$P(W) = (p(\text{female}), p(\text{predator})), \quad Q_s = \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}, \quad Q_r = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}, \quad m = 2 \tag{6.2}$$

A stable strategy pair for a similar game is depicted in figure 6.2.[1] Because both female and predator send signal s_1 in this model, its functional content is $< 1,\ 0 >$. Treating functional content as truth-evaluable, predator flashes are therefore false.

What about the statistical content of the firefly flash? Flashes are sometimes given by females, and sometimes by predators. There is a positive correlation between the signal and both states. Therefore, the statistical content has positive values in both parts of the vector: $< \log \frac{p(\text{female}|\text{signal})}{p(\text{female})},\ \log \frac{p(\text{predator}|\text{signal})}{p(\text{predator})} >$. This indicates that the probability of both female and predator is increased upon sending of the signal, though the probability of female is increased more (on the assumption that females are more prevalent and/or better at accurately

[1] Shea et al. (2017, p. 17) take the game to represent sexual signalling in which the two senders are high-type and low-type males. In our case, the receiver payoff upon responding favourably to a predator signal should perhaps be less than zero. However, I have kept the payoffs the same for ease of comparison. Certainly, if predators became too prevalent in the population (if $p(\text{predator})$ became too high) the strategy pair depicted in figure 6.2 might no longer be stable.

sending the signal than predators).

Interestingly, female signals correlate with predators and predator signals correlate with females. Statistical content describes all the possible states that co-occur with a signal. Functional content is motivated by the need for explanatory work of a different kind. It says *why* signals continue to be sent. Functional content in the firefly system refers to females only, because it is only females whose signals promote system-beneficial behaviour.

Although Shea et al. (2017) do not commit to the theory, functional content is a way to formalise sender-receiver teleosemantics. Functional content captures the same explanatory role for content that teleosemantics aims to explicate. When a signal brings mutual benefit due to a combined effect of the receiver's act and some external world state, that signal bears an explanatorily relevant relation to that state. This is at least implicit in scientific theorising about biological signalling, and functional content offers to formalise those explanatory practices.

Finally, it is worth noting that functional content can be defined such that the same signal possesses two different functional content vectors, associated with sender and receiver. Recall, vector components of a signal are zero when it fails to prompt a greater-than-baseline payoff. But some signals may cause behaviours that bring greater-than-baseline payoff to the receiver but not the sender. Then we can distinguish two functional content vectors: one associated with the sender, that contains zero for this particular vector component; and one associated with the receiver, for which this vector component has a positive value. Distinguishing these vectors may help elucidate selective forces that act differentially on agents in evolutionary interactions. Senders are more motivated to send signals with positive functional content vectors. Indeed, individualised vectors would quantify, in some sense, the

motivation for each agent in using the signal. Especially in cases where interests diverge, different states will play a stabilising role for sender and receiver strategies. In the firefly example, all functional content vectors have entries that are either zero or 1. The definition of functional content therefore formalises Millikan's account for these signals: all signals have a unique content, and deceptive signals gain false content as a result of mimicking co-operative signals. The power of the functional content definition lies in the ability to capture mixed contents; cases in which more than one vector entry is positive, and so upon being normalised these entries take values other than zero and 1. In this sense, the account might be considered a generalisation of teleosemantics: it goes beyond signals with precise content and offers to capture vague cases.

6.3.3 Artiga's options correspond to modelling choices

Artiga considers the relative merits of rejecting MIMICRY and rejecting INCOMPATIBILITY. In the context of signalling games – especially in light of functional content – rejecting INCOMPATIBILITY is the better option. Recall there are two ways to pursue this option. Either predator sending-mechanisms are categorised as the same type as female sending-mechanisms, or predator signals are categorised as the same type as female signals. These two options correspond to two different ways of modelling the firefly system. Categorising predator sending-mechanisms as the same type as female sending mechanisms corresponds to modelling senders as a single type with partial common interest with receivers, as in model 6.2 above. Alternatively, categorising predator signals as the same type as female signals corresponds to modelling senders as two different types, one of which has full common interest with the receiver, the other of which has fully conflictual interests with the receiver.

The first model is depicted in figure 6.2. Predator and female signals have the same functional content. Signals are true when sent by females and false when sent by predators. This is probably the canonical way to model the situation. Skyrms (2010, §6) discusses alternative representations, but asserts that this option captures the deceptive, contentful aspect of the predator signal (even though he works with statistical rather than functional content).

Consider the second option, on which two different types of sender can send the same type of signal. We can even model these as different signals that are indistinguishable for the receiver.[2] Now signals have different functional content when produced by different senders. Female signals have the usual functional content. Predator signals have no functional content (or at least a vector with only zero entries), because no receiver act brings greater-than-baseline payoff for both parties simultaneously. Predator signals would, however, have an individualised sender-functional content that was nonzero: there is an act the receiver can perform that brings the sender greater-than-baseline payoff.

These proposed solutions provide a tentative means for retaining and justifying the biological practice of explaining behaviour by reference to the content of mimicked signals. But there is a deeper justification for attributing content in cases of divergent common interest. *A generalised form of communication theory encompasses costly signalling theory. Our teleosemantic analysis of content in signalling systems generalises too.* This claim is not in tension with the work just described. It can instead be read as supporting aspects of their accounts from a different direction. To begin, we will establish a core model from which three different theories can be drawn out.

[2]Realistically, predator signals are somewhat inaccurate. The more accurate they become, the more selection moves receivers to more discerning perceptual abilities. This in turn prompts better mimicry, and the cycle continues. More sophisticated signalling games would surely provide insight into these dynamics.

6.4 A core model

Sign-reading behaviour trades off metabolic expenditure for error. Understanding the evolutionary logic of sign-reading behaviour entails figuring out how to convert both error and expenditure into fitness costs. Sign-reading is stable if there is a net benefit.

Signalling behaviour trades off two sets of metabolic costs (one each for sender and receiver) for two sets of error costs (likewise). Understanding the evolutionary logic of signalling behaviour entails figuring out how to convert both error and expenditure into fitness costs. Signalling is stable if both parties achieve net benefit compared to what would result from non-signalling strategies.

Error can be interpreted in terms of proper function, and proper function entails a history of selection. Typically, errors motivate a change in behaviour.[3] Selection processes that sustain signalling behaviour are sensitive to error. For evolutionary agents the selection process is natural selection. For learning agents it is trial-and-improvement, or whatever learning procedure they employ. For cultural agents it might be imitation. For engineered systems it is whatever the engineer considers a fault, including not performing up to the specification. In each case, errors destabilise current behaviour by inducing selective forces that serve as a corrective for error-prone strategies.

Our core model includes two agents with payoffs determining behaviour. They are causally connected by an intermediary signal, but they need not cooperate (figure 6.3). Selection pressure on each derives from two sets of costs. *Metabolic expenditure* is paid by

[3] An error may not in fact *trigger* a change in behaviour: it may be too minor for the system to 'notice'. But errors must at least contribute to outcomes that selection could in principle be sensitive to; otherwise, there is no clear sense in which they are errors. Also, I am not claiming that errors are a *necessary* condition on behavioural change due to selection, though they might be.

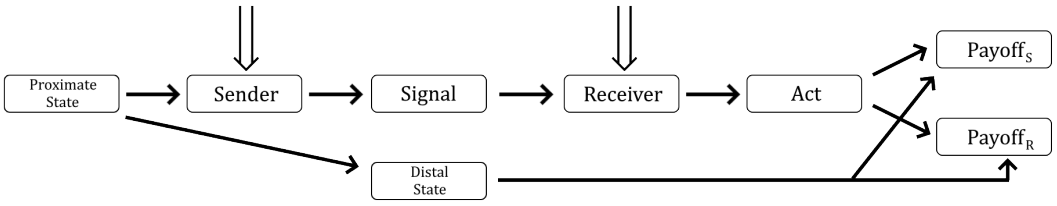

Figure 6.3: The core model, a generalised version of the basic sender-receiver model. The sender observes a proximate state and sends a signal. The receiver observes the signal and performs an act. The act, together with a distal state, determines payoffs for sender and receiver. Because the payoffs of sender and receiver are different, there is no guarantee of cooperation. Single arrows denote causal relationships. Double arrows denote selection processes.

an agent when they send or receive a signal, observe a state, or perform an act. *Error costs* are derived from the familiar payoff matrices. In a sense, an error matrix – more properly a **distortion matrix** – is an 'inverted' payoff matrix. Just as payoff matrices determine the benefit due an agent given a combination of state and act, distortion matrices determine the penalty for each state-act pair (see section 6.4.2 below for more).

Figures 6.3-6.6 depict the core model and three versions of it, corresponding to our three target theories. In the **core model** (figure 6.3), a distal state together with the receiver's act determines payoffs. The sender can observe a proximate state (which may be identical with the distal state). The signal completes a causal loop between the distal state and the payoffs. The greater the correlation between signal and distal state, the greater control the receiver has over the payoffs. Double arrows denote selection processes of any kind (evolutionary, ontogenetic, cultural, rational, etc.).

In **sender-receiver teleosemantics** (figure 6.4), sender and receiver are selected to cooperate. The distal state is the descriptive correctness condition of the signal. The proximate

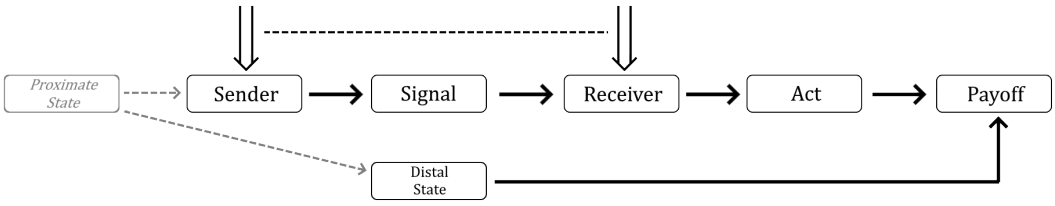

Figure 6.4: Sender-receiver teleosemantics. Sender and receiver cooperate fully, meaning there is just one set of payoffs (represented by a single variable). This is also represented by a dashed line between the double arrows, which are selection processes acting on sender and receiver. Selection acts on both sender and receiver equally. The proximate state is greyed out because sender-receiver teleosemantics emphasises the distal state instead.

state need not be causally related to the distal state, though for non-accidental proper functioning they must at least be correlated. Contrast *informational teleosemantics* (Neander, 2017), according to which the proximate state is the signal's content. Again, double arrows stand for any selection process that induces proper functions.

In **costly signalling theory** (figure 6.5), sender and receiver payoffs differ, so they are not necessarily selected to cooperate. The proximate (behaviour-causing) and distal (payoff-relevant) states are usually both features of the sender itself. For example, the payoff-relevant state may be quality or willingness to fight. Metabolic expenditure ensures a correlation is maintained between this feature and the signal, making the signal reliable (see below for more). Double arrows denote natural selection or rational choice.

Finally, in **rate-distortion theory** (figure 6.6), the payoff-relevant state is identical to the proximate state. The receiver's act is an attempt to reconstruct the proximate state, with payoff increasing as it reconstructs the state more accurately.[4] Double arrows denote an en-

[4]In fact, this sentence and the last are only true in traditional applications. Rate-distortion theory applies to more interesting situations too. First, the CEO problem (Berger et al., 1996) (El Gamal and Kim, 2011, §12.4) is the problem of reconstructing a source given several noisy observations of it. This roughly fits the schematic

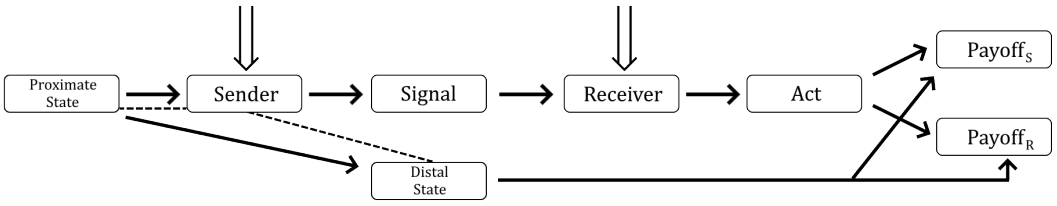

Figure 6.5: Costly signalling theory. Both the proximate state that prompts sender behaviour, and the distal state that contributes to payoffs, are usually considered to be physical properties of the sender (depicted by dashed lines connecting all three variables). Payoffs can diverge, leading to deceptive signalling.

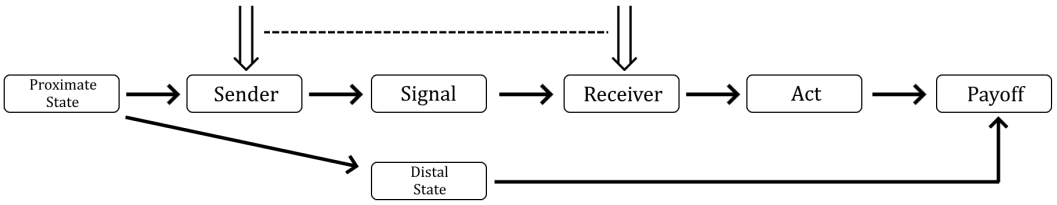

Figure 6.6: Rate-distortion theory. The proximate state is a string of symbols, and the distal state is (usually) identical to it. The receiver's act is an attempted reconstruction of this string, and payoff is determined by a measure of its accuracy.

gineer's rational choice about how sender and receiver should encode and decode the signal.

Let us now explore costly signalling theory and rate-distortion theory, and their relation to teleosemantics. All three theories employ versions of the core model, each for a different purpose. Costly signalling theory explains signalling behaviour in the face of conflicting interests. Rate-distortion theory describes optimal transmission rates for perfectly cooperative systems with nonzero tolerable error. Teleosemantics defines correctness conditions for signals.

in figure 6.6, but with several nodes in place of the proximate state and arrows pointing into them from the distal state instead of the other way round. Second, as will become crucial below, the distortion function itself can measure the correctness of any functional behaviour with an appropriate payoff distribution. Thanks to Manolo Martínez for these details.

6.4.1 Costly signalling theory

A scenario familiar to biologists is that of signalling under the threat of deception (Maynard Smith and Harper, 2003, §2), (Searcy and Nowicki, 2005, §4) (Bruce et al., 2017; Bruner et al., 2017). Behavioural ecologists typically describe the problem of animal signalling as the problem of explaining why signals remain honest. The same is true in other kingdoms too. Bacterial colonies often comprise groups with divergent interests. Signalling helps bacteria coordinate behaviour such as resource sharing. When members of signalling partnerships have diverse evolutionary interests, cooperative communication is under threat of dishonesty (Bruce et al., 2017). Dishonesty would lead to system breakdown. Nevertheless, in animals, bacteria and elsewhere, signalling is rife in the biological world.

Costly signalling theory is a game-theoretic approach to explaining stability of signalling systems despite incentives to deceive. In a sentence, it explains how metabolic expenditure can keep signals honest. Consider a predator-prey interaction like that of *Anolis* lizards and their snake predators (Leal, 1999). Snakes would be better off not engaging with strong lizards because they might lose the fight (and their dinner). Lizards would be better off acting so as to cause snakes to assess them as strong. This evolutionary logic leads to the lizard's pushup display, indicating endurance capacity and the ability to escape an attack. Such displays have been shown to correlate positively with fight-relevant features, recommending a costly signalling analysis (Leal, 1999) (see case study).

Pursuit deterrence: the pushup displays of the *Anolis cristatellus* lizard

Male lizards of the species *Anolis cristatellus* have a lot to contend with. Not only do they face predation from the snake *Alsophis portoricensis*, but they must compete with male conspecifics for the attention of females. Fortunately, they have evolved a signalling strategy that serves both functions. When faced with a predator or rival male, the lizard bobs its head up and down repeatedly in what has been described as a 'pushup' motion. These displays take a significant amount of effort to pull off. Because that effort could be otherwise expended in resisting predation or fighting a love rival, it has been suggested that the pushup is a costly signal. Leal (1999) tested this hypothesis in the case of predation with favourable results. He showed that the number of pushups performed in the presence of a model snake positively correlates with the endurance capacity of the lizard. Together with the assumption that snakes assess the endurance capacity of their prey, this suggests that the pushup display is a costly signal. The metabolic resources required to produce the display increase proportionally with the underlying feature of the sender the display is designed to indicate.

Debate over costly signalling theory usually concerns such issues as how to assign costs, how to measure them in real populations, alternative explanations and how-possibly scenarios produced with the aid of models. Maynard Smith and Harper (2003) discuss various ways in which signals can be kept honest, advising researchers not to assume metabolic expenditure is always the correct mechanism. Searcy and Nowicki (2005) treat metabolic expenditure (the "handicap mechanism") as the most prominent mechanism by which sig-

nals remain reliable, but consider alternatives too. Fraser (2012) disputes the taxonomies of both Maynard Smith and Harper (2003) and Searcy and Nowicki (2005). Fraser offers a new taxonomy of costs, arguing that separating imposed from intrinsic costs offers a more fruitful analysis than that of Searcy and Nowicki (2005). Bruner et al. (2017) highlight the difference between evolutionary cost and metabolic expenditure, demonstrating that in certain cases the link between these two is non-obvious, leading to consequences that differ from traditional theoretical results. For this reason, I continue to refer to metabolic cost as expenditure. 'Metabolic cost' will mean the ultimate fitness costs induced by metabolic expenditure; 'cost' will be used when the distinction between expenditure and cost is irrelevant.

Instead of further elaborating costly signalling theory, I want to look at the very simple explanatory approach that underlies Grafen's seminal mathematical treatment. Appendix D presents the mathematics; here I give an informal summary.[5] Grafen (1990) described a model in which senders are incentivised to exaggerate their quality, as occurs in sexual signalling from males to females. Males make a decision how expensive a signal to send, where all are in principle capable of sending any signal. Females observe the signal and assess the quality of the male on the basis of it. Since low-type males are incentivised to signal at greater intensity and enjoy the rewards of higher female assessment, there must be some mechanism preventing this from happening.

Grafen showed that under certain conditions, if signalling is stable then signals are honest and costly. Signal cost is the mechanism that maintains signal honesty. By assumption, signalling is stable, which means receivers must be associating signals of a certain intensity with males of a certain quality. It does not pay low-type males to send higher-intensity

[5]This summary corresponds to the *major handicap result* in appendix D.

signals, because to do so they must expend more resources than a favourable assessment can recoup. As a result, the way receivers associate signal intensity with male quality is a stable strategy too. All males are signalling in a way that corresponds to their quality, so females can take them at their word.

Importantly, this explanation is strategic. Signal expense incentivises senders to make strategic decisions that are revealing – honest – to receivers. After Grafen, others such as Bergstrom and Lachmann (1998) have shown that expenditure is not necessary to maintain honesty. What is important, however, is that signal expense *can* play a role in explaining a certain kind of stability.

When biologists use game theory models to explain behaviour, they often search for mechanisms that maintain **evolutionarily stable strategies** (ESS). An ESS is a strategy which, when adopted by the entire population, cannot be supplanted by a low-frequency mutant due to selection alone. In this sense the strategy is stable: it will not be out-competed by any alternative strategy. Grafen showed that strategic costs are one mechanism by which signalling under divergence of interests remains stable.

Costly signalling can be compared to the core model (figure 6.3). Sender and receiver do not have perfectly aligned interests, so selection is not joint selection. That is why there is no dashed line between the double arrows in figure 6.5. The proximate state on which the sender conditions their behaviour and the distal state that determines payoffs are typically very closely related, or even identical, in the costly signalling model. The idea is usually that sender quality (distal state) is determined by the sender's genes, and sender behaviour is under genetic control. In this case, both the proximate and distal states may be controlled by a third variable, upstream of both, that represents genetic elements. Grafen's strategic

explanation captures the success of genes that ensure quality-appropriate signalling.

In sum, costly signalling theory provides explanatory tools for understanding stable communication between agents with divergent interests. It takes the core model (figure 6.3), finds instances of it in nature (typically, those that match figure 6.5), and offers an explanation for why these systems remain stable over time. Metabolic expenditure is not a universal stabilising mechanism, however. Some signalling systems are kept stable by other means. The core model supports many explanations why and how a system remains stable. The costly signalling model specialises in a single kind of answer to this question.

6.4.2 Rate-distortion theory and capacity-cost theory

Although many philosophers of biology will be familiar with costly signalling theory, few philosophers of any stripe will be familiar with rate-distortion theory. That is a shame, because a lack of appreciation for the formal tools available within communication theory has led to pessimism about its usefulness for naturalistic approaches to meaning.

Shannon's fundamental results, about achievable transmission rates over noiseless and noisy channels, hold in ideal cases. They assume that the goal is error-free message reconstruction. Engineers quickly realised that perfect accuracy is not always sought. A signalling system may tolerate some error in the reconstructed message. A more general theory of communication therefore allows for signalling at a lower rate, incurring tolerable error.

Reflecting on this, Shannon and others developed a generalised form of communication theory. Augmenting his earlier mathematical apparatus, Shannon (1959) introduced a measure of the cost of inaccuracy. When faced with a given communication problem, engineers can choose an acceptable error cost and design coding methods that do not exceed that cost

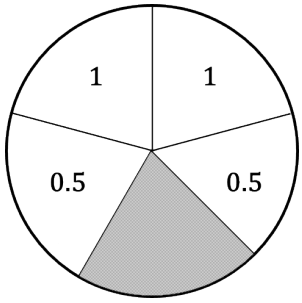

Figure 6.7: A rate-distortion problem. A wheel is divided into five segments. The receiver must choose the correct segment (shaded). They are penalised for choosing an incorrect segment, but the penalty is lower for segments closer to the correct one. The example is from Shannon (1959, p. 327).

on average. Furthermore, the goal of communication need not be reconstruction of a symbol string. The formal apparatus of the theory requires only well-defined events and a probability distribution over them. An example will make these points clear.

Suppose you want to transmit the position of a wheel. The wheel can be in one of five positions, with each equiprobable. The receiver may choose the incorrect wheel position, and both sender and receiver incur a cost. Suppose the two positions adjacent to the actual state incur a small cost, while the other two incur a large cost (figure 6.7). The accuracy of the wheel can be improved by expending more metabolic resources in the fight against noise. Reducing noise entails increasing transmission rate. In general, we want to know the minimum transmission rate (i.e. minimum metabolic resources expended) that incurs at most a specified amount of error.

The resulting extension to communication theory is called rate-distortion theory. **Distortion** refers to the tolerable error, which sets a lower bound on transmission rate. If you can afford to lose accuracy, you can save on transmission resources. The appropriate distor-

tion function for a given task will be largely determined by context. Conversely, a maximum available rate sets a lower bound on distortion. The lesson of the theory is that given a goal, a channel, and a maximum average cost, a minimum rate exists that achieves the goal without exceeding the cost.

Costs range from zero (all outputs correct) to a maximum determined by the receiver's best possible guessing strategy. In the wheel example, a blind receiver's average distortion is $\left(\frac{1}{5}\times 0\right)+\left(\frac{1}{5}\times\frac{1}{2}\right)+\left(\frac{1}{5}\times\frac{1}{2}\right)+\left(\frac{1}{5}\times 1\right)+\left(\frac{1}{5}\times 1\right)=\frac{3}{5}$. So when $\frac{3}{5}$ average distortion is acceptable, the minimum transmission rate is zero: you do not need to transmit in order to meet that level of performance. Maximum transmission rate, at which distortion is zero, is $\log 5$ bits/signal: this is what is required to distinguish one out of five equiprobable states.

Blahut (1972) and Arimoto (1972) independently derived an optimisation algorithm that delivers the appropriate minimum rate for every value of allowed average distortion. Alternatively, given a maximum available rate, the algorithm delivers the minimum achievable cost. Rate-distortion diagrams (figure 6.8) can be read in either of these two ways.

So far we have talked about how error changes with rate, but there is another significant cost to bear: the cost of transmission. Metabolic resources are required to power transmission, no matter what medium is being used. In an engineering context these are typically electronic messages, so resources include the energy and time required to send the signal, the material used to build and maintain equipment by which transmissions are sent and received, and wires across which the signal is sent. In general, expending metabolic resources increases the rate of signalling.

Just as before, there are two ways to approach this trade-off. If you have a goal, a channel, and a *maximum* average metabolic expenditure you can sustain, there is an optimisation algo-

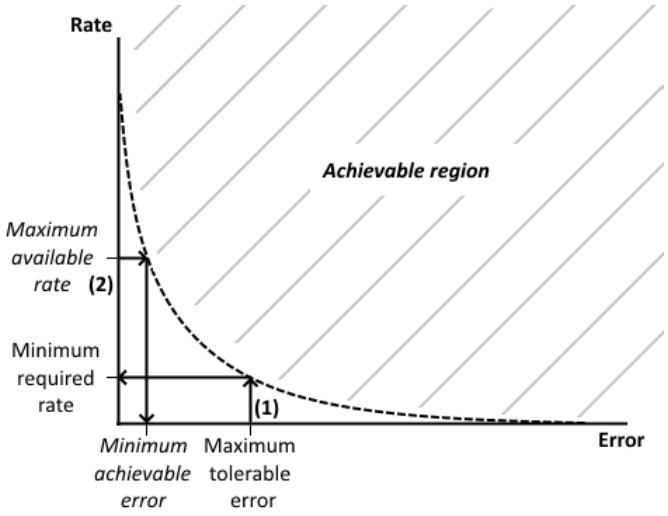

Figure 6.8: A rate-distortion curve. The algorithm that produces the curve can be used to solve either of two problems. (1) Given a maximum tolerable error cost, the algorithm finds the minimum required rate. (2) Given a maximum available rate, the algorithm finds the minimum achievable error.

rithm that tells you the *maximum* rate at which you can transmit. Alternatively, if you have a *minimum* rate at which you *must* transmit, the algorithm tells you the minimum metabolic expense you can hope to incur. The subdiscipline of communication theory that deals with this trade-off is called **capacity-cost theory**.[6] Jimbo and Kunisawa (1979) presented an algorithm that is a counterpart to the one given by Blahut and Arimoto for rate-distortion theory. de Polavieja (2002) applies a similar algorithm to a capacity-cost problem in a biological context (see appendix E).

Historically rate-distortion theory and capacity-cost theory have been kept apart. The first considers how to mitigate error by increasing transmission rate (or conversely, how

[6]The term 'capacity' is used here because the maximum achievable rate when there is no expenditure constraint is the capacity of the channel.

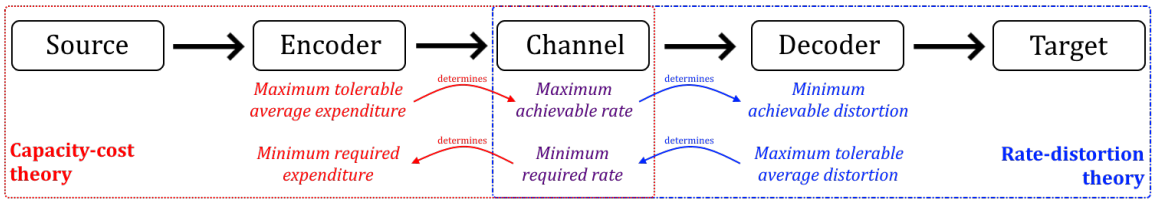

Figure 6.9: Together, metabolic expenditure and distortion requirements place upper and lower bounds on transmission rate. If you want to minimise distortion, you are constrained by expenditure; if you want to minimise expenditure you are constrained by distortion. Transmission rate acts as an exchange rate.

to get away with lower transmission rate by accepting greater error). The second considers how to mitigate metabolic expenditure by *decreasing* transmission rate (or conversely, how to get away with higher transmission rate by accepting greater metabolic expenditure). Supposing these two approaches could be combined, the result would be a trade-off of two different kinds of cost. On one hand, if you want to decrease error by increasing transmission rate, you must expend metabolic resources to do so. On the other, if you want to decrease metabolic expenditure by decreasing transmission rate, you must accept increased error. The best signalling strategy would embody the optimal trade-off between error and metabolic expenditure. Transmission acts as a kind of exchange rate in this calculation, describing how much error you can offset by expending metabolic resources (figure 6.9). Then, depending how error and expenditure compare, you can determine the optimal rate at which to transmit.

In some cases the two costs may be incommensurate. Indeed, that seems to be the reason why the two approaches are kept distinct in engineering contexts. A maximum allowable error is chosen and the transmission rate is adjusted to ensure the system stays below that level of error. The question how to choose a maximum tolerable error (and the question

how to measure the *cost* of error as a function of error) is answered differently depending on context. Available metabolic resources act as an independent constraint.

In subsequent sections I explore the link between these aspects of communication theory, their relationship to costly signalling theory, and how the resulting theoretical edifice supports a teleosemantic approach to meaning. First, let us briefly recap those aspects of teleosemantics relevant to the question at hand.

6.4.3 Sender-receiver teleosemantics

Teleosemantics is a theory of signal content. Signals are intermediaries between coadapted senders and receivers, that guide the receiver in performing a jointly beneficial function (Millikan, 1984, p.96ff) (Millikan, 2004b, §6) (section 1.2). In bringing mutual benefit to sender and receiver, the receiver's behaviour stabilises the signalling relationship. The signal guides this task, hence contributes to the stable recurrence of this relationship in future.

Signal content enters the picture when we consider how the signal successfully guides the receiver. Of a range of actions receivers could perform, conditioning their behaviour on signals leads them to greater success. Signals improve receiver success because they bear appropriate relations to relevant parts of the world.

Teleosemantics defines content in terms of mapping relations between signals and their satisfaction conditions (figure 6.4). Focusing on descriptive relations, a signal's correctness condition is that state of affairs that must obtain in order for the receiver's behaviour upon receipt of the signal to be nonaccidentally successful (figure 1.4, page 18). In short, a signal maps onto a state of affairs and this mapping helps the receiver perform its job successfully.

Mapping relations explain how signalling behaviour persists through selection.[7] The title of this chapter loosely indicates these two features. That a signal maps onto the world – that it has 'sense'[8] – explains receiver success. How that success contributes to the signal's own proliferation – to its stable future production – is the causal effect the signal has on the receiver that helps bring about a mutually beneficial outcome.

Teleosemantics faces a problem dealing with realistic cases because it defines meaning with respect to perfect common interest. It deals only with functions that are codesigned. Senders and receivers perform their functions by cooperating. Their interests do not diverge. However, many examples of biological signals operate between agents whose evolutionary interests are not perfectly aligned. Even within individual organisms, signals may be sent between components that do not cooperate. Different parts of a body have the same ultimate interests but may disagree over the best way to satisfy them. Bodies are rowdy parliaments of criss-crossing intentions. Teleosemantics presents an ideal that biology fails to live up to.

6.5 Unifying the theories

I want to argue that unifying costly signalling theory and rate-distortion theory supports teleosemantics, so I must first show how they are unified.

There is a provocative similarity between dishonesty and noise. Speaking about dishonesty in costly signalling scenarios, Grafen says:

[7] That is, once they have originated and become established as part of a signalling system, mapping relations help explain successful signalling behaviour. See section 5.3 for an argument that teleosemantics does not fall foul of an objection on the basis of circular explanation.

[8] *Not* Fregean sense. Not 'sense' in any theoretically loaded way. Just a loose synonym for the mapping relations defined by teleosemantics.

> Cheats impose a kind of tax on the meaning of the signal.
>
> Grafen (1990, p. 535)

On the other hand, speaking about noise in cooperative scenarios, Shannon says:

> Nature takes payment by requiring just that much uncertainty, so that we are not actually getting any more than [the capacity] through [the channel] correctly.
>
> Shannon (1948b, p. 410)

This is more than just an idle analogy. From the perspective of a cooperating sender-receiver pair, dishonest senders are a kind of noise. Furthermore, the difference between environmental noise and dishonesty-as-noise mirrors the distinction between efficacy costs and strategic costs. Let us explore these ideas in more detail.

6.5.1 Communication theory encompasses costly signalling

Intuitively, reduced common interest means reduced transmission rate. Effort expended in signalling must be paid back by receiver behaviour, but reduced common interest means receivers are less incentivised to perform behaviours rewarding for senders. Conversely, conditioning behaviour on received signals only pays when the signal corresponds to some outcome-relevant world affair. But reduced common interest means senders are less incentivised to produce signals that correspond to world affairs according to shared mapping rules. As common interest decreases, so too does the incentive to signal. Relating this point to our earlier idea that common interest is a kind of exchange rate between sender and receiver costs, the more interests diverge the more prohibitive the exchange rate becomes.

We already know that communication-theoretic codes are a special case of game-theoretic strategies (section 3.2). de Polavieja (2002) and Martínez (2019) assert an even tighter link. Martínez shows how to interpret costly signalling as a rate-distortion problem. The idea is

roughly as follows. Distortion was conceived as a measure of how detrimental it is to reconstruct a symbol incorrectly. As a measure of the severity of error, it penalises the entire system. In the central model there is no sense in which a given error might be good for the sender and simultaneously bad for the receiver. Engineered communication systems have unified interests by design: error for one is error for all. Not only is there no deception in such systems, there is no sense in which there could be deception.

These two aspects of the distortion measure – that it measures pairs of symbols and is shared by sender and receiver – can be generalised. I have already explored extensions of communication theory beyond the symbolic paradigm. Distortion need not be a measure over symbols but can be a measure over outcomes more generally. And instead of d for the system we can define d_S for senders and d_R for receivers. Distortion need not be a system-level measure but can apply to individual agents. In this form, distortion starts to look a lot like payoff matrices from game theory models. Payoff matrices define how good are different acts in different situations, distortion matrices say how *bad* they are. Indeed, Martínez (2019, p. 5) shows how to convert payoff matrices into distortion matrices and vice versa (see appendix B). In this way, game-theoretic models of costly signalling interactions are recast as communication-theoretic models of rate-distortion trade-offs.

6.5.2 Dishonesty as noise

Here is how we can interpret this situation. For the sender, the receiver is a constraint. For the receiver, the sender is a constraint. From the point of view of the *signal*, imperfect common interest of sender and receiver is a constraint. Martínez (2019, p. 1) makes a similar point, using the phrase "channel-first perspective" to denote the signalling system as a jointly

constructed entity with its own evolutionary interests.

In section 6.4 we saw that engineering applications of rate-distortion theory and capacity-cost theory are usually kept apart. In the biological case, on the other hand, the costs of communication error and metabolic expenditure can be made commensurate by considering their impact on fitness cost. Indeed, costly signalling theory is predicated on a trade-off between error and expenditure. In Grafen's model, error for a sender just means failing to attract a receiver (while error for a receiver means accepting low-type senders, foregoing the chance to mate with high types). Senders pay costs to attract receivers more reliably – that is, to reduce the probability of the 'error' of failing to mate.

de Polavieja (2002) suggests we can interpret dishonest senders as noise impacting the cooperative system between honest senders and receivers. In cooperative systems, transmission rate is increased (and metabolic costs paid) in order to combat noise. The introduction of dishonest senders (increasing noise) incentivises the honest senders to boost their signals (increasing metabolic expenditure) and make receivers reliably respond (decrease error). Although the central aim of costly signalling theory is to describe conditions on stability, it achieves this by imagining optimal prudential decisions made by signallers. This explanatory tactic highlights the link with communication theory.

de Polavieja's perspective highlights two kinds of metabolic expenditure, corresponding to the two types of signal design identified by Guilford and Dawkins:

> [T]he first, strategic, component of signal design is what determines whether or why (in fitness terms) the signal receiver responds appropriately, whereas the second, efficacy, component affects the probability that the signal, once given, will reach its target destination and elicit a response at all.
>
> Guilford and Dawkins (1991, p. 2)

Strategic costs are paid to overcome receiver reticence, while efficacy costs are paid to overcome environmental noise (see sections 1.4 and 7.2). The term 'metabolic expenditure' conflates both kinds of cost to encourage the idea that from the sender's perspective receiver reticence just is a kind of noise. It is a state of affairs external to the sender that hinders the signal's proper function. In the same way that the signal/cue distinction collapses when what is being explained is receiver behaviour (section 1.3), the signal/influence distinction collapses (and with it the strategic/efficacy distinction) when what is being explained is sender behaviour.

One way of interpreting the costly signalling trade-off is to say *since error destabilises the system, paying to increase the rate helps stabilise it.* Prevalent dishonesty threatens to overwhelm receivers until they no longer respond. The system evolves to a point at which signals are no longer heeded, and soon they are no longer sent. This is what it means for error to destabilise the system. The rate-distortion perspective highlights that error and noise both destabilise signalling systems. Even when cooperating, senders must expend metabolic resources to allow receivers to hear their signals over the noise. In other words, even when strategic costs are not necessary to ensure stable honesty (as in the case discussed by Bergstrom and Lachmann (1998)), efficacy costs are required to ensure signals are sufficiently noise-free to be worth responding to.

From a receiver's perspective, when we shift between game theory and communication theory, we can treat dishonest signallers as a source of noise. Either way, what we have is a force running counter to the well-functioning of the system.

6.6 A stability principle

Error has two sources: dishonesty and noise. In both behavioural ecology and communication theory, error destabilises signalling systems. In both cases, increasing metabolic expenditure helps overcome error, hence contributes to system stability. This concept is a general idea that applies in engineered and biological systems. An engineered system is 'unstable' to the extent that it does not yet satisfy an engineer's error criterion. The engineer will then induce metabolic expense – changing the system – until it satisfies their intent. At this point they no longer change the system and it is stable. The role of selection is played by the engineer's hand. This link between error and system stability is a good candidate for a fundamental principle of communication. Call it the **stability principle**:

> Stability principle: a signal is erroneous to the extent that it contributes to outcomes that would destabilise the communication system of which it is a part.

Certainly there are problems with this proposal as stated. For example, consider the ESS pair in the deception game, depicted in figure 6.2. Treating the game as a representation of the firefly case, one might reasonably argue that predators do not strictly destabilise the system by sending the mimicked flash signal. Predators are sufficiently rare (or send the signal sufficiently rarely), relative to payoffs, that their false signals do not perturb the system from its stable state. That is what it means for this pair of strategies to be evolutionarily stable, after all. Although each system has a certain level of fault tolerance, fault can still be characterised as a destabilising force. I am not concerned with formalising the stability principle.[9] I am more interested in exploring how an idea reasonably similar to it would

[9] Martínez (2015), responding to Godfrey-Smith's suggestion that deception could be defined in terms of "non-maintaining signals" (Godfrey-Smith, 2011, p. 1295), gives reason to think precisifying the principle adequately will be rather difficult: there are "games in which deceptive signals can be arbitrarily frequent, without

accord with teleosemantics. To move forward, I must set aside worries about the specific form of the claim.

6.6.1 Teleosemantics treats the stability principle as definitional of correctness conditions

The stability principle refers to signals being erroneous. This implies there is an antecedently defined mapping between signals and world affairs. And indeed in both formal theories this is so. In rate-distortion theory, as elsewhere in communications engineering, signal meaning is determined by the code (see sections 3.2 and 5.5). Codes are defined by the engineer who chooses sender and receiver strategies simultaneously. Costly signalling takes place in a game theory setting, in which signal meaning is determined by the joint behaviour of sender and receiver. These behaviours are selected by coevolution of successful strategies. Signallers come to agree on the meaning of a signal, and this agreement ensures system stability. In both cases, correctness conditions are stipulated, and a link with stability is then discovered.

Teleosemantics suggests a small but significant shift in perspective. Sometimes in mathematics we can interpret a theorem as a definition of one of the terms within it (Soare, 1996, §3.4). For example, suppose you have an intuitive concept of the shape that the word 'rectangle' refers to. The concept does not come with necessary and sufficient conditions attached, but it does distinguish paradigmatic rectangles from marginal cases, and can help identify non-rectangles. Suppose you offer the following theorem:

---

this undermining information transfer between sender and receiver" (Martínez, 2015, p. 215). Intuitively we would expect all deceptive signals to count as erroneous, so in formalising the stability principle we may be forced to either give up the intuition or refine the principle further.

> Theorem: a quadrilateral is a rectangle if and only if it is a parallelogram with diagonals of equal length.

In seeking to prove the theorem, you might check paradigmatic rectangles for the properties the theorem claims they necessarily have. But while interrogating your concept RECTANGLE, you might find that it is either too broad, too narrow, too vague or informal, or unsatisfactory in some other way. You might therefore decide to sharpen your concept by *defining* rectangles this way:

> Definition: a quadrilateral is a **rectangle** if and only if it is a parallelogram with diagonals of equal length.

By making this move, you sharpen the intuitive concept of rectangle. Some aspects of the intuitive concept might be lost, but the benefits of rigorous definition are obvious. This is conceptual engineering, and it drives much mathematical progress.[10]

Consider now the case at hand. Intuitive notions of truth and falsity are acceptable for rate-distortion and costly signalling theories because workers in the respective fields agree how to apply those concepts in their explanations. The stability principle operationalises a notion of error and draws a link between error and stability:

> Principle: a signal is erroneous to the extent that it contributes to outcomes that would destabilise the communication system of which it is a part.

Faced with this principle, a philosopher might mistakenly believe that formal theorists have a robust definition of correctness and error in hand. I suggest instead that we do not need

[10] See for example Woodward on the concept of continuity, or Tao on the concept of measure (Woodward, 2003, p. 8) (Tao, 2013, §1.1). Note that these are mathematical concepts. Conceptual engineering in mathematics is geared towards mathematical goals, rather than those of philosophers or the folk. Nonetheless, philosophers can learn from this method. The concept of causal specificity might be a good case (Griffiths et al., 2015) (see footnote 19, page 46).

an antecedent definition of correctness and error to make sense of the principle. We can and should interpret the stability principle as a definition:

> Definition: a signal is **erroneous** to the extent that it contributes to outcomes that would destabilise the communication system of which it is a part.

And the definition of correctness conditions follows.

On the rate-distortion perspective, one source of error is noise. It seems rather bizarre to talk of noise being 'true' or 'false'. Even when talking of signals that would be correct but for noisy interference, we would tend to assert that they are 'distorted' rather than false. It is more natural to say that *falsity is distortion at the point of origin*. A signal is false when, at the time the sender produces it, it already fails to map. This never happens in standard models of communication theory. But it happens a lot in costly signalling theory.

Defining correctness conditions in terms of system stability is broadly equivalent to asserting sender-receiver teleosemantics (Millikan, 2004b, §6), (Artiga, 2016b). Recall how teleosemantics defines a signal's (descriptive) correctness condition: the state of affairs that must obtain in order for the receiver's behaviour upon receipt of the signal to be nonaccidentally successful. The signal's truth condition is the state of affairs that accounts for the receiver's behaviour being successful. Signals mapping correctly onto their truth conditions selects for responsive receivers, which in turn selects for compatible senders. Mapping onto truth conditions is a matter of degree.

A few comments on the relationship between teleosemantics and the other two theories are in order. First, at least one teleosemanticist has recognised the two fundamental costs described here. Speaking of signals in evolutionary systems, Millikan (2004b, pp. 71–2) says, "They need only to succeed often enough to offset their own production costs (energy

and resources used) plus any negative effects resulting directly from failures." Here "failure" means failing to produce causal effects that contribute to evolutionary success of the system. "Negative effects resulting directly from failures" means, more or less, error costs. Second, teleosemantics describes the core model of communication theory. This point has been established elsewhere (chapter 3), (Artiga, 2016b, p. 495) (Martínez, 2019, p. 1) so I will not dwell on it.

6.6.2 Error and proper function

Above I said that the term 'erroneous' seems to indicate a preexisting attribution of content to signals. But at its most basic, an error measure is simply a measure of the cost of inopportune behaviour. In game theory models, the error measure is simply a difference in payoffs. In biological contexts it is the fitness cost. The correctness of signals derives from the correctness of behaviour. The latter is formally enshrined in mathematical theory, though open questions remain concerning application and interpretation.

Pursuing this line further, if distortion is employed as a measure of error, but error is determined by proper function performance, then distortion is in general a measure of proper function performance. This fits well with Martínez's point that a distortion matrix is just an inverted payoff matrix (Martínez, 2019, §3). Distortion matrices represent (mathematical) functions from an ordered pair of input alphabet and output alphabet to the real numbers, whereas payoff matrices represent (mathematical) functions from an ordered pair of states and acts to the real numbers. It is clear by now that alphabets generalise to states and acts (section 3.2). It should also be clear that a payoff matrix defines a measure of proper functioning: if you do well, you get rewarded, and it is those rewards that determine functional

behaviour.

6.6.3 Functional content formalises the stability principle

Functional content is one way to formalise the stability principle. It assigns values to vector components when the corresponding state helps contribute to continued sending of that signal in that state. For example, in the bottleneck game (model 6.1) there is a sense in which signal s_1 induces a 'faulty' action when sent in w_2: it prompts act a_1 when a_2 would have been better. In restricted scenarios like this, senders cannot help but send 'false' signals some of the time. But in situations where interests diverge, false signals may become more prevalent. Then they harm the well-functioning of the entire system, and threaten to reduce transmission rate to zero.

Shea et al. (2017) do not mention destabilising uses of signals, and they do not define functional content for signals sent out of equilibrium. Nonetheless, because functional content picks out those states with which a signal must coincide in order to contribute to stability of the system, it also picks out states in which that signal would *de*stabilise the system. Continuing the bottleneck example, each time s_1 is sent in w_2 it reduces the receiver's incentive to pay attention to the signal. When responding to this signal in this state, it would do just as well acting without paying attention to the signal. It may be objected that although such uses of a signal are non-stabilising, that does not entail they are de-stabilising. However, in very many real-life cases, there will be some metabolic cost to responding to a signal over ignoring it. When signalling does not pay, selection will often favour devices that economise by refusing to signal or respond.

Furthermore, metabolic costs can play a role in shaping functional content. For example,

Shea et al's figure 2 (p.12) depicts a situation in which the functional content of a signal depends on the relative frequency with which subsequent acts are performed. These proportions will be shaped by both kinds of cost: if one is more metabolically costly or provides less net benefit, it will be performed less often or not at all.

Finally, recall that different vectors can be associated with the same signal, capturing the difference between states that stabilise sender behaviour and those that stabilise receiver behaviour. More work needs to be done to add to this formal repertoire.[11]

There may be other ways to formalise the stability principle. In particular, different formulations may be required for evolutionary models driven by dynamics other than the replicator equation. On the evidence so far, prospects are good for formalising teleosemantics in a manner that attributes content in conflictual scenarios. I now turn to an objection.

6.7 An objection

Objection. Defining error in terms of system instability puts the cart before the horse. A definition of error is required in order to explain failure, hence instability. Receivers sometimes fail *because* signals do not map onto the states they ought to. Error cannot be defined in terms of failure on pain of circularity.

To further press the objection, consider a distinction between two kinds of reason for receiver failure. On one hand, a receiver might fail to perform its function because the signal maps incorrectly. For example, a disoriented bee might dance inaccurately. On the other hand, a receiver might fail to perform its function *despite* the signal mapping correctly.

[11] I note that Millikan has indicated acceptance of the idea that signal content is sometimes ambiguous in this way; see the conversation between Millikan and Sterelny at Millikan (2018b, 1h07m50s-1h11m20s).

For example, a worker bee flying towards a patch of flowers indicated to it by a scout's waggle dance might be preyed upon by a bird. Such features are not typically represented in signalling games, but in real life there are many obstacles to achieving a goal. In order to distinguish these kinds of cases, there needs to be a mapping between signal and world that can be evaluated independently of the receiver's subsequent performance. The proposed definition fails to provide it.

Response. My response is rather flat-footed. Before there is a convention, there really is no difference between signal failure and failure for other reasons.

Consider a situation where sender and receiver have common interests but happen to be at a stage in their evolution where they perform incompatible strategies. In such a case, should we say that the sender is sending wrong or the receiver is receiving wrong? I submit that neither can consistently be asserted. There is as yet no convention, so no mapping to speak of. Signals in this system do not yet have determinate content. Only given an agreed signalling strategy can signals go wrong in respect of it.

Take for example the very first 2x2 game in *Signals* (Skyrms, 2010, p.10ff.) (model B.1 in appendix B). At the start, the population is 50/50. At the end, one strategy wins out. Perhaps with hindsight we could say 'this half of the population were wrong, and *that is why they died out*'. But you could not say that ahead of time.

Perhaps we can relate this to Birch's suggested refinement of Skyrms's account of propositional content (Birch, 2014b). On Birch's account, the content of a signal is what it would be at the nearest separating equilibrium. In the above example, signals that are exactly equidistant between two equilibria have no content. Perhaps transmission rate could be used as a measure of how far a system is from equilibrium. The transmission rate in the

initial state of Skyrms's model is zero (because $H(X) = H(X|Y) = 1$), and this is the furthest possible point from equilibrium. This is a good reason to treat signals as contentless at that point.

Bringing this back to the natural case, before there has been selection for the waggle dance, there really is no difference between failing to find food and being eaten by a bird. Selection over millions of years smooths out all those unusual circumstances, so the right Normal explanation picks out the descriptive content of the signal: *the state of affairs common to all those occasions*; that is, the presence of food in a location spatiotemporally related to the form of the dance.

Rejoinder. My interlocutor may reject as too radical the idea that signals continuously change their meanings as senders and receivers change strategies. Such a view renders meaning too cheap and ephemeral, and threatens our understanding of semantic content as (say) a determinate mathematical function from possible worlds to extensions.

Response to rejoinder. But such a view is at odds with what we know of *linguistic* meaning, let alone less paradigmatic kinds of content. Words really are continuously changing their meanings, as speakers and hearers update their conversational strategies for a wide range of reasons. The same is even true for signal form, whether linguistic or biological, regardless of their meaning. The point to bear in mind is that the meaning of a signal in a well-functioning communication system remains stable *enough* to support non-accidental joint function. That's how it is with human language, and if linguistic meaning is the paradigmatic example of semantic content, we ought not be concerned that other bearers of meaning display similar changeability.

6.8 Conclusion

A unified picture of the fundamental nature of communication is emerging. Sense and stability go hand in hand. Appealing to stable strategies to explain or ground standards of correctness has been hinted at in recent naturalist philosophy (Birch, 2014b) (Sterelny and Fraser, 2017) (Dennett, 2017, §3). These moves would benefit from recognising the relationship between stability and proper function. Formalising teleosemantics in terms of functional content (or similar) renders it readily applicable to cases of imperfect common interest. Prospects are good.

Chapter 7

Evolutionary communication theory

7.1 Introduction

Biologists explain signalling behaviour by appealing to core tenets of evolutionary communication theory. In this chapter I outline in as much detail as possible what I take that theory to be.

Evolutionary communication theory is an explanatory paradigm. It assumes that the goal of communication is coordinated function, and that transmission is hindered by noise, divergent interests, and physiological constraints. All of these considerations are part of an extension to the original theory of communication. Biologists draw on this paradigm without employing a univocal name for it. That is the job for which I have designed the label 'evolutionary communication theory'.

In section 7.2 I motivate the account by outlining important features of biological signals not explicitly captured by classical communication theory. Then, in section 7.3 I characterise evolutionary communication theory as an explanatory paradigm that appeals to coordinated function, divergent interests and physiological constraints. The next few sections survey existing work that implicitly (section 7.4) or explicitly (section 7.5) uses explanatory concepts from communication theory. Finally, I respond to objections (section 7.6).

7.2 Background and motivation: the information debate in behavioural ecology

Parts of this section are taken from Mann (2018).

In section 2.2 we saw how to interpret informational measurements in a paradigmatically cooperative signalling system. Several reasons counted in favour of using the honeybee waggle dance as a case study. Through decades of research there is plenty of data available on bee dances and their effect on compatriot bees. Strong eusociality entails a negligible threat of free riding or defection. It is reasonable to assume signalling is fully cooperative, avoiding complex problems of cross-purpose and deception. But the question must be faced: what becomes of this methodology when few or none of these assumptions hold?

Although we dealt with some objections to the use of informational concepts in behavioural ecology in section 2.4, there remain disanalogies between communication theory and animal signalling. This section motivates evolutionary communication theory by considering challenges from Dawkins and Krebs (1978) and several joint and individual papers from Owren, Rendall and Ryan (hereafter ORR) (Owren et al., 2010; Rendall and Owren, 2013; Rendall et al., 2009; Ryan, 2013).

7.2.1 Three problems with the informational approach: content, constraint, and coadaptation

As a consequence of the abstraction required for informational description, questions of behavioural and developmental constraints have often been ignored. Animals do not have great freedom of action, and natural selection does not have free rein in shaping their behaviour.

Many purported instances of signalling can be shown to be distinctly manipulative (Rendall et al., 2009, p. 237) (Owren et al., 2010, pp. 766–7). Consider two populations of senders and receivers without perfectly aligned interests, such as males and females of a single species. Suppose, as a general rule, overt behaviour can adapt faster than the cognitive architecture underlying perception. Then receivers will often lag behind in the perceptual arms race. Senders will be quick, on an evolutionary scale, to exploit perceptual biases while receivers will be slow to rectify them. Receiver responses are then determined by trade-offs between the need to respond to relevant perceptual stimuli without being duped too harshly.[1]

This story of trade-offs goes beyond signals that are 'about' something, such as male quality, and includes behaviours that cannot be evaluated in terms of honesty. ORR point out that many examples of what we typically think of as signals are not (or only degenerately) contentful, because their function is tied specifically to receiver affect or attention. For example, males of some fish species display coloration similar to that of their female receivers' prey, the point being to catch a female's attention (Owren et al., 2010, p. 767). Importantly, being better at attention-grabbing need not correlate with quality. The reason for this male adaptation is not a corresponding female adaptation; it is a prior female adaptation for an entirely different function. The following case study is an example.

[1]There are at least three reasons to doubt the universality of this scenario. First, Bergstrom and Lachmann (2003) describe conditions under which the slower-evolving organism enjoys the benefit, a phenomenon they call the Red King effect. Second, Brusse and Bruner (2017) argue that in cooperative cases it benefits both sender and receiver that senders evolve faster than receivers. Third, once we consider associative learning on individual timescales, the balance of power in the arms race shifts: receivers learn quickly to distrust false signals, but senders are slow to develop new production mechanisms.

Surfperch sensory drive

Surfperch are fish of the family *Embiotocidae*. They live among kelp forests off the coast of California, where dappled light produces a high variance in underwater microhabitats. Related species of surfperch have diverged evolutionarily, in part due to the different visual properties of the environments they inhabit. Different species vary with respect to which of two properties of visible light they are optimised to detect: chromaticity and luminance.

The **chromatic** component of a light source is its colour. It is detected via *inhibitory* connections in visual processing, downstream of retinal photoreceptors. In contrast, **luminance** is the brightness of a light source. It is detected via *excitatory* connections in visual processing. Because these two properties are detected by contradictory processes, there is an evolutionary trade-off in sensory systems designed to detect them both. In particular, since surfperch are dichromats (they have only two types of retinal cone), improvement in chromatic detection almost inevitably entails degradation of luminance detection and vice versa (Cummings, 2007, p. 531).

One would expect surfperch sensory systems to respond to this trade-off by favouring whichever of the two features was most ecologically relevant. Whether chromaticity or luminance is most important depends in large part on the visual properties of the background environment. In environments with highly variable luminosity, chromatic detection is favoured. It would be reasonable for fish in such environments to have sensory systems adapted to detect colours rather than luminance. Indeed, that is what Cummings (2007) reports. She first infers a common ancestor of five

> surfperch species, showing that this ancestor would have a sensory system balanced between chromatic and luminosity sensitivity. Of five extant species studied, two displayed a shift towards a chromatic-favouring sensory system (*Hypsurus caryi* and *Micrometrus aurora*). These two species occupy habitats with higher luminance variability (Cummings, 2007, p. 540). In contrast, the three species in more stable habitats displayed a shift towards favouring luminance detection (*Embiotoca jacksoni, E. lateralis, Damalichthys vacca*). Cummings rules out the possibility that relatedness between species explains this variation.
>
> Furthermore, male surfperch develop colouration designed to get the attention of females. Different species display different colours, and these features correspond to the visual sensitivities of females. The best explanation for these correlated features is that sensory systems adapted to the aforementioned trade-off, then signalling colouration followed suit (Cummings, 2007, p. 542).

According to Cummings' hypothesis, male surfperch colouration is an example of **efficacy design** rather than strategic design (Guilford and Dawkins, 1991). Efficacy design comprises features of signals that improve their *detection*, whereas strategic design refers to features of signals that affect the receiver's *assessment* of the sender. Sensory-driven colouration gets the attention of females, but it need not indicate anything about the underlying quality of males.

It can be difficult to differentiate these two kinds of design in part because different hypotheses sometimes entail the same predictions. For example, on the present hypothesis, "female preferences for orange or red coloration [entails *efficacy design* due to a] visual

foraging bias for carotenoid-rich food" (Cummings, 2007, p. 543), but on an alternative hypothesis "this preference is also predicted by indirect or direct benefit models if these colors serve as indicators of male carotenoid-foraging ability or condition [entailing *strategic design*]" (Cummings, 2007, p. 543). In short, it might be the case that male colouration evolved in part due to what it indicates about male quality. But, crucially, it might not. What is important is that efficacy considerations may be sufficient to explain signal design.[2]

ORR want to draw on cases like the surfperch to argue that information is a bad concept to employ in animal signalling. There are three lines of argument they can draw on. First, traits and behaviours designed as attention-grabbing appear not to be *contentful*. Where male colouration does not indicate male quality, there is nothing that the signal is 'about', hence it ought not be described as carrying information. Second, physiological *constraints* play a larger role in shaping the form of the signal than is usually afforded by information optimists. Idealised models taken from game- and communication theory assume that signallers can employ any strategy. The trade-offs that shaped female surfperch sensory systems demonstrate that receiver strategies are much more constrained than those models assume. Third, male coloration is not a consequence of *coadaptation*. It is an adaptation that exploits pre-existing sensory processes in females. Call these the problems of *content*, *constraint* and *coadaptation*. I will take them one by one.

Consider first the problem of contentless signals. In the example, fish coloration is not contentful, because it does not correlate with anything. ORR argue that in situations like these, "courtship signals do not evolve to carry information about signaler quality" (Owren

[2]Presumably there is a 'baseline' level of efficacy design for primarily strategic signals: a signal must be at least detectable in order to cause a receiver's response.

et al., 2010, p. 767). However, this does not count against a communication-theoretic perspective on animal signals. Consider what we concluded from the surfperch study. Paying careful attention to the distinction between efficacy and strategic design, Cummings pointed out that male surfperch colouration might not be a case of strategic design – that is, they might not have evolved to indicate male quality. But their efficacy features are just as much part of a communication-theoretic approach as their strategic features. Indeed, ORR emphasise that courtship displays often show features designed to allow them to overcome background noise (Owren et al., 2010, p. 766). But this is precisely the problem that channel coding, in its evolutionary guise as efficacy design, solves.

Furthermore, emphasising the distinction between efficacy and strategic design does not entail that the latter is *never* responsible for signal evolution. ORR's claims are consistent with the existence of signals that evolved to indicate features of the sender or the wider world. Indeed, a given signal may have evolved for reasons to do with efficacy *and* strategic design, each ingredient being important to varying degrees. It does not follow from the fact that there are signals whose design is not strategic that there are *no* signals whose design *is* strategic. It is the latter kind of signal that is paradigmatically contentful.

Finally, the objection from contentless signals contains a hidden assumption: that in order to be contentful, signals must *indicate* a state of the world. To the contrary, signals can also contain instructions how to act. This could include an instruction to *pay attention!*, opening the channel for further signalling. (Instructive signals may also be construed as having a rather degenerate descriptive aspect, such as *something here!*.) Of course, whether or not a descriptive aspect is attributed, the instructive aspect still requires coadaptation. By assumption the fish in the example do not meet this condition. I deal with this wider problem

below, but the point to take away here is that instructive signals need not have the familiar indicative content usually assumed by paradigms like costly signalling theory.

ORR's second line of attack against the informational approach concerns physiological constraints. The informational approach seems to assume that all conceivable variants of a given signalling behaviour are available for evolution to select among. Then the explanation for (say) a given receiver response is its adaptive benefit. But signalling behaviours whose form is primarily attributable to physiological and developmental constraints are not apt for this kind of explanation. In the case of surfperch, females respond to male colouration not because of an evolutionary benefit it brings, but because their visual processes are constrained in their accuracy. In this case, constraints are imposed by a trade-off between chromatic and luminance sensitivity. Different kinds of sensory systems will face different kinds of trade-offs, and are subject to constraints as a result. Canonical signalling explanations draw on receiver function to explain signal form. Developmental constraints crowd out this explanation.

To respond, note that signal form can be determined by multiple evolutionary and developmental forces at once. Signal form and receiver behaviour can be constrained, and these constraints can play a role in adaptive explanation. There is nothing inconsistent in the idea that signals nonetheless bear mapping relations to world affairs. Animal signals span a broad volume within the interaction hypercube (section 1.5). Constraints and content both *contribute to*, rather than *compete for*, signalling explanations. For example, in the case of the waggle dance, certain physiological constraints have been proposed to explain a systematic error in dance performance (Preece and Beekman, 2014) (see the case study on page 282). This does not threaten the mapping relation between the form of the

dance and the world affair it signifies. It simply reduces the accuracy, hence the efficacy, of the dance. Indeed, the extent to which the dance is hampered by this constraint could be measured in terms of how it affects transmission rate. Such measures could contribute to phylogenetic models of the origin of the dance (see section 7.3 for suggestions how to represent constraints in sender-receiver models, and section 7.4 on the work of Dumont et al. (2014) testing hypotheses about evolutionary optima).

Now, it is of course true that communication theory does not typically consider constraints on signalling behaviour. The central model assumes a free choice of code. There are costs to sending longer signals (corresponding to power costs in electrical equipment), and perhaps signal detection theory discusses issues related to costs of precise recovery of signals in the face of noise. What is needed here is a principled extension to communication theory, one that considers optimal joint behaviour from the perspective of restricted design capabilities. Models of communication from an evolutionary perspective would afford much less freedom of choice of encoding. In particular, the code – the form of the signal – that develops through coevolution will be simultaneously constrained by competing adaptive needs of both sender and receiver. Crucially, it could still be the case that coadaptation plays a role in explaining signal form, even though several different kinds of constraint play a role too. As with contentless signals, the prevalence in nature of constrained signals does not threaten the propriety of an informational approach to biological communication.

Finally, consider ‘signals’ that are not the result of coadaptation. The informational paradigm in behavioural ecology assumes that the function of signals is to transmit information between sender and receiver. But in situations where sender and receiver have conflicting interests, there is no such function for signals to serve. Animals without common

interest do not perform joint functions. Since the informational approach defines signals in terms of joint function, it would be forced to conclude that these interactions are not signals. This is at odds with the practice of behavioural ecology. The majority of animal signals are threatened with divergent interests. To argue that such behaviours are not truly signals would be to narrow the application of that term so radically as to render it almost useless. Thus, ORR and others promote a more inclusive definition of signal in terms of the influence one animal exerts over another.

One option here is to bite the bullet. The 'signals' highlighted by ORR as part of their critique do seem to be influence, the third category of interaction we discussed in chapter 1. Just as with cues, influence blends into signalling when differential benefit becomes coadaptive. Just because there exist interactions that predominantly benefit senders, does not mean there are no cooperative signals in nature. It certainly does not entail that cooperative signals should be understood in terms other than information. Moreover, only absolutely conflictual interactions lack a joint function, cases in which the entities have no common interest whatsoever. Although the majority of animal signals are threatened with divergent interests, they are also maintained by common interests. Sexual signalling can continue even in the face of deceptive males, because it is at least sometimes beneficial for females to mate. There are borderline cases, and the features of signals that are due to cooperation rather than conflict are still apt for informational explanation. Finally, influential behaviours sometimes display similar design to signals. Repetition of stereotyped action patterns, and high visual salience of morphological features, are common aspects of efficacy design that work both in cooperative and conflictual contexts. As a result, the design principles of evolutionary communication theory may help us understand the form of influential behaviours just as they

do cooperative signalling.

The tripartite categorisation of signals, cues and influence is a rough covering of many possibilities. Existing models within evolutionary game theory are explicitly designed to study these situations. In particular, work inspired by Skyrms's sender-receiver framework has explored the limits of communication in noncooperative settings (Martínez and Godfrey-Smith, 2016; Wagner, 2012), as well as different rates of evolution (Brusse and Bruner, 2017). Further, making this distinction will likely help ORR achieve at least one of their goals. They aim to stop theorists searching for the wrong explanations of signalling behaviour. By distinguishing three categories of animal interaction, they can vividly point out that the majority of cases are closer to influence than coadapted signals. The hypercube analysis presented in section 1.5 allows us to distinguish cooperative situations, typified by the social insects, from those with multiple conflicting evolutionary forces. Above we saw that distinguishing efficacy and strategic design allows practitioners to discover the correct explanation of signal form. The same holds here: distinguishing signals, cues and influence allows practitioners to determine into which category a particular interaction falls. Intermediate cases exist. But these categories remain valid.

7.2.2 Evolutionary communication theory accommodates a wider class of interactions

The problems just discussed motivated Dawkins and Krebs's original suggested redefinition of signalling:

> To summarize the point of view we are adopting: as an inevitable byproduct of the fact that animals are selected to respond to their environment in ways that

> are on average beneficial to themselves, other animals can be selected to subvert this responsiveness for their own benefit. This is communication.
>
> Dawkins and Krebs (1978, p. 285)

The authors explicitly conflate influence and coadapted signals. By contrast, the informational approach highlights an important difference between them. Just as parasites must walk a fine line between exploiting and destroying their hosts, so senders must strike a balance between reliably manipulating receivers and driving them out of existence. We proceed on the assumption that signalling can only exist when some mechanism helps maintain equilibrium, preventing one or both parties going extinct. One obvious example is common interest, when the same behaviour brings benefit to both sender and receiver. But there may be other mechanisms maintaining the stability of an interaction, thus helping to explain why we observe it.[3]

Theoretical and modelling work should continue to expand their horizons regarding manipulation and constraint. One party in an interaction characterised by partial common interest may be at risk of extinction, but which one (and why) will be an empirical question in each case. The mechanisms evolution has thrown up that prevent extinction, thus prolonging signalling behaviour, may be many and varied. Part of the work of theory-building is to find a taxonomy of such mechanisms and outline their symptoms, which can then act as diagnostics for field workers to employ (see for example Hurd and Enquist (2005)). Dawkins and Krebs promote the view, later picked up by ORR, that in the vast majority of cases we will find general-purpose receiver perceptual mechanisms being exploited by special-purpose sig-

[3] One question I do not address is the extent to which the stability of a phenomenon explains our observation of it. Since evolution by natural selection is plausibly still causing changes in populations, we would expect to observe at least some unstable behaviours. What we need, I suppose, is a theory that delivers predictions about both change and stasis – even if the timescale of such predictions is too large for us to easily test.

	Interactions (general)	*Signalling (specific)*
Normative framework; Rational agents	Game theory	Communication theory
Descriptive framework; Evolutionary agents	Evolutionary game theory	**Evolutionary communication theory**

Table 7.1: Evolutionary communication theory in relation to three well-known theories.

nalling behaviour. So be it: theoretical work will uncover a wider class of forces than those manifested in nature.

All of this is apt – none of it threatens our account of information. Evolutionary game theory has been used to explore interactions of divergent interests, and it shades into communication theory when the interests are common and there are fewer physiological constraints. Evolutionary communication theory borders these two theories (see table 7.1). Weakening the assumption of common interest does not preclude mathematical analysis. It means that the mathematics gets more general and its interpretation perhaps more difficult. It means mechanisms other than 'honesty' maintain signalling interactions. It means we should look not just for static equilibrium behaviour, but ongoing arms races. ORR, and others such as Sarkar (2013), believe that none of this mathematics will have anything to do with communication theory. Throughout this thesis, I have tried to show that is an overly narrow conception of the links between communication theory and game theory.

Let us end this section on a positive note. ORR have clearly presented important considerations for evolutionary perspectives on communication. Animal interactions are beset by divergent interests and physiological constraints. Natural selection does not give rise to perfect forms. Evolutionary models tend to make assumptions trading realism for tractability. We must incorporate a diverse range of signalling phenomena, such as male surfperch

colouration, into our investigations.

7.3 What is evolutionary communication theory?

Evolutionary communication theory frames explanations of signalling behaviour in terms of four principles:

1. The goal of communication (conceived as a piece of behaviour, a morphological structure, or an emission) is coordinated function.
2. Environmental noise threatens the efficacy of signalling behaviour.
3. Divergent interests threaten the stability of signalling behaviour; signals become more cue-like or influence-like as interests diverge.
4. Physiological constraints narrow the range of available signalling behaviours.

Incorporating all four principles into a single framework suggests a way to organise biological signalling research. Philosophers such as Skyrms (2010, §3) and Scarantino (2015) downplay the distinction between signals and cues. Biologists such as Dawkins and Krebs (1978) and ORR downplay the distinction between signals and influence. I want to emphasise both distinctions, and keep them at the heart of evolutionary communication theory.[4]

The first five chapters concerned the first two principles, while chapter 6 investigated the third. A few remarks on physiological constraints are therefore in order.

[4]It should be noted that the account does not *define* signals in terms of information. As Scott-Phillips (2008) points out, a definition of signal based on coadaptation (such as that given in chapter 1) already entails that signals carry information.

7.3.1 Physiological constraints

Constraints reduce the space of strategies available to a biological device. By correctly identifying constraints, a scientist can reduce the space of adaptive explanations they can appeal to.

There will likely be different communicational constraints for senders and receivers. Receivers face trade-offs because their sensory systems are typically multi-purpose. They must deal with cues as well as signals. Often they must reliably recognise salient objects across a range of environments.[5] The surfperch case study provided a particularly vivid example of a sensory trade-off. Inherent properties of the dichromatic visual system constrained available strategy space. By identifying this constraint, Cummings was able to suggest and test for a hypothesis about divergence of surfperch species from an inferred ancestral state. Senders, on the other hand, face different kinds of constraint. Take for example the systematic error in the honeybee waggle dance, described in the following case study.

Systematic error in the waggle dance

As we saw in section 2.2, the honeybee waggle dance indicates the direction and distance of food. In performing the dance, sender bees are subject to a peculiar kind of systematic error. The directional component of the dance is relatively less accurate for food sources that are closer to the hive (Preece and Beekman, 2014). Now, one would expect that accuracy in general decreases for food sources that are further away from the hive, and this is indeed what happens when distances become very great. But within an ecologically relevant region scouted by bees, the following curious error

[5] Online learning helps: once the capacity to recognise is sensitive to environmental variation within a single lifetime, one-size-fits-all sensory systems are less of a liability.

occurs. The area of the region indicated by the dance remains roughly the same at different distances. This is unusual: if signal accuracy were uniform over different distances, the area of the indicated region should increase as distance from the hive increases. Suppose for example the dance carries 3 bits of information about the angular location of food, meaning it reduces the space of possible directions receivers should fly down to $\frac{1}{8}$ of its original size. The further a food source is from the hive, the larger the circle around which directional information is provided, hence the larger is the angular region of search space remaining once it has been cut down to $\frac{1}{8}$. Therefore, if angular information in the dance was constant, search space would increase for food patches further from the hive. But that is not what is observed. Surprisingly, search space remains constant for a range of distances from the hive, meaning that *more* angular information is provided by the dance for food patches that are further away.

Initially, biologists believed that the tendency for search area to remain constant was an adaptive feature of the waggle dance. Various suggestions about the evolutionary benefit of such an 'error' were advanced. For example, the sender bee might itself be uncertain about the location of the food patch it found, thus it directs receivers to search in that general area. Another adaptive hypothesis states that food patches are more likely to be found near to each other than at random, meaning that receivers who miss the indicated food source are nonetheless likelier to find food in its vicinity than if they had searched the whole region.

However, arguments that systematic error is an adaptive feature of the dance do not withstand scrutiny. Preece and Beekman (2014) adduce several lines of evidence in

favour of an alternative hypothesis. They convincingly argue that sender bees are constrained in how precisely they can convey angular information. The constraint is tighter for food patches closer to the hive, leading to less accurate indication.

What kind of physiological features could account for this phenomenon? Recall that distance from the hive is indicated by the length of the straight run of the dance, while direction is indicated by the angle off the vertical (figure 2.1). It turns out that it is difficult for a sender to find the correct angle off the vertical each time it returns for another straight pass. Like a drunk motorist walking a policeman's white line, it helps to build up some momentum. For short runs, bees do not have time to align themselves with the correct angle. On average, short dances will have a larger angular error. On the other hand, when the straight run is long, initial wobbles are smoothed out and the average angular precision is higher. It is a simple consequence of the physical relationship between two aspects of the dance that angular error increases as food is closer to the hive.

Selection can only choose from the available variants; this will always be a smaller set than the conceivable variants.

7.3.2 Modelling constraints in signalling games

The main difficulty theorising about constraints is determining what they are in the first place. If there are difficulties in modelling constraints, they are less important. However, I note that these kinds of constraints are not typically included in models inspired by Skyrms (2010). Since I am working in that tradition, this should be understood as a call for more appreciation

of the role of constraints in these models, as well as more innovative methods of representing constraints within models generally. Most optimistically, a thorough understanding of how constraints can be manifested within models could illuminate what constraints there are on real organisms. (By comparison, rock-paper-scissors dynamics were initially investigated in models by Maynard Smith, before later being discovered in the wild. Here we are talking about constraints rather than dynamics, but the idea is comparable: studying features of models sometimes makes us more aware of different possibilities in the real world.) How might we model constraints in sender-receiver games?

First, the space of strategies may be artificially constrained by simply omitting certain pure strategies from the model. For example, we might demand that senders be unable to send the same signal in different states, and so strategies $P(S|W)_3$ and $P(S|W)_4$ in model B.1 (appendix B) are unavailable. For a more complex example, consider the receiver strategies depicted in model 7.1. This is a four-signal game in which receiver $P(A|S)_1$ can distinguish signals s_1 and s_2 reliably but is unable to distinguish signals s_3 and s_4 reliably. Conversely, $P(A|S)_2$ can distinguish signals s_3 and s_4 reliably but is unable to distinguish signals s_1 and s_2 reliably. Suppose s_1 and s_2 represent signals distinguished by luminosity, and s_3 and s_4 represent signals distinguished by chromaticity. The model can then be used to represent the surfperch case study on page 271.

$$\underbrace{P(W) = \left(\frac{1}{2}, \frac{1}{2}\right)}_{\text{States equiprobable}} \quad \underbrace{Q_s = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad Q_r = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\text{Perfect common interest}} \quad \underbrace{m = 4}_{\text{4 signals available}}$$

$$P(A|S)_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad P(A|S)_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{7.1}$$

Second, we may add a 'genetic' component in order to model developmental constraints. Instead of explicitly constructing strategy matrices, we could allow a program to generate each matrix stochastically based on certain parameters. Instead of passing strategies directly to the next generation, the program along with its parameters is inherited. In such a model, developmental constraints could be represented by certain features of the developmental program. Situations like that depicted in model 7.1 can be faithfully ensured by a genetic program that contains intrinsic restrictions on the receiver strategy matrices it produces. For example, the program may have a fixed amount of 'discrimination ability' to apportion between conditional acts, corresponding to constraints on developmental resources available to the surfperch. The program would be able to produce a strategy that could distinguish exactly two signals perfectly (e.g. S1 and S2). In addition, it could tell when a signal was neither of its two favoured types (e.g. that it is either S3 *or* S4), though it could not tell which of the two remaining signals it was (cannot distinguish S3 *from* S4). This discrimination capacity is $\log 3$ bits: it can in effect discriminate three signal types: S1, S2 and S3-or-S4. So the developmental program would be allowed $\log 3$ bits of discrimination capacity, to appor-

tion among its different conditional acts. Starting with such a constraint and a fixed sender population that adopts either the S1-vs-S2 strategy or S3-vs-S4 strategy, a selective process might converge on the receiver strategies depicted in model 7.1. Two distinct lineages could emerge, mirroring the two surfperch strategies recorded by Cummings (2007).

Third, we may attempt to go beyond the traditional framework and consider structured signals. Consider for example a signal comprised of two components, distance d and angle θ that jointly indicate a location from an origin. This can be achieved by splitting the state into two states corresponding to distance and angle, and giving the sender two corresponding strategy matrices. We might constrain the signals that can be emitted, such that shorter distances entail less precise angles. This would be one way to model the constraint on the honeybee waggle dance described in the case study on page 282. The best way to implement this constraint might be to include a noise function, such that θ is perturbed by a magnitude randomly chosen from a distribution whose shape and size is a consequence of d.

Fourth and finally, de Polavieja's model in appendix E assigns differential costs to signals. Optimal signalling requires balancing these 'metabolic' costs with the need to transmit information at a useful rate (see also chapter 6). de Polavieja (2002) intends his model to represent neural signalling, where the relevant metabolic costs include consumption of the energy-carrying molecule adenosine triphosphate (ATP). Limits on available resources clearly constitute a constraint on signalling behaviour.

7.3.3 Optimality in biology

A fully worked out mathematical framework underpinning evolutionary communication theory could provide general conditions on which signalling should be selected for. Such a

result would be akin to a complicated version of Hamilton's rule (Birch, 2014a; Hamilton, 1964a,b).

Explanatory frameworks that employ optimality models can be useful even when the predictions of those models fail. Predication/observation mismatch sometimes points to hidden costs or constraints, generating new testable hypotheses and suggesting observations (Sterelny, 2013). Furthermore, as Birch (2017, §2.6.2) points out in the context of Hamilton's rule, such frameworks enable researchers to identify when a particular phenomenon *cannot* be explained by recourse to selection mechanisms. For example, suppose two creatures are observed to communicate, incurring low metabolic costs and/or low error, despite apparently divergent interests and no identifiable physiological constraints binding them to the interaction. In such a case the resources of the framework are exhausted, and researchers must look elsewhere for an explanation. For Hamilton's rule, Birch (2017, p. 60) calls explanations "(partially) non-selective" when they must appeal to evolutionary processes beyond natural selection to explain proliferation of a trait within a population. In the present context, partially non-selective explanations must appeal to at least one other evolutionary process to explain persistence of signalling behaviour. It is as though the combined forces of positive selection are not sufficient to outweigh interfering factors, so something else must explain why the organisms in question reliably communicate.[6]

Conceiving of evolved traits as optimal solutions to ecological problems may strike some as illicit. Gould and Lewontin (1979, p. 581) famously derided adaptationist "faith in the power of natural selection as an optimizing agent". However, as Parker and Maynard Smith

[6]In such a case, if selection is not the source of communicative behaviour, then that behaviour is not strictly proper-functional, and so not really communicative at all. However, behaviour that *seems* functional demands explanation, even – especially – when it is not.

(1990, p. 27) point out, optimality models "serve to improve our understanding about adaptations, rather than to demonstrate that natural selection produces optimal solutions." Using optimality models to investigate evolved forms does not require assuming that present traits are optimal. The practice may only require assuming that traits, if they are adaptations, have moved in a certain direction in design space in part because of the improved performance that movement entails. The following case study is an example.

Adaptive radiation in leaf-nosed bats

Recent work on adaptive radiation combines phylogenetic reconstruction with optimisation to explain the variety of forms within clades. Morphological systems afford different performance optima, corresponding to appropriate forms for different environmental niches. For example, in lizards, longer limbs can increase speed, but shorter limbs can improve stability on narrow substrates (Vanhooydonck et al., 2006, p. 4516). Dumont et al. (2014) investigated which of two performance criteria was most important in the selection pressures underlying the cranial shape of different species of leaf-nosed bat. The performance criteria were *mechanical advantage*, a proxy for bite force, and *von Mises stress*, a proxy for structural strength (where low stress correlates with high strength). Using a morphological space whose dimensions are cranial length and width, the authors generated a *morpho-performance space* in which computer-generated cranial shapes are evaluated based on their performance against the two criteria. They used plausible phylogenies calculated from genetic data to constrain the historical trajectories of different clades across morphospace from the inferred ancestral state. They then compared two sets of models: one in which selection occurs with

respect to optimal cranial shapes that differ based on the dietary regimes of different species, and the other in which selection does not occur and variation in cranial shape is due to a random walk across morphospace.

The study found that mechanical advantage was a more significant factor in explaining adaptive radiation. Short-faced bats, for example, exhibit high mechanical advantage because their diet of hard canopy fruits requires a high bite force In contrast, nectarivores display low mechanical advantage because their long, narrow snouts – a requirement for long, nectar-reaching tongues – do not support high bite force. This approach brings quantitative methods to previously informal hypotheses about adaptive radiation and speciation, such as Darwin's iconic finches (Marsh, 2015).

Notably, at least one of the study's predictions failed. The authors predicted that frugivores and Short-faced bats would exhibit selection for low von Mises stress, on the grounds that they need higher bite forces as part of their diet, and this requires physical support from greater structural strength. One explanation for this lack of selection is that "the cranium never approaches the critical stress (i.e., the material strength of bone) during normal activities, and therefore routine stress is unlikely to play a role in shaping the evolution of cranial form" (Dumont et al., 2014, p. 1445).

What is important is not the precision of measurements that led to this conclusion, but the possibility of quantifying performance criteria. This is required before optimality models can be applied. Perhaps future work will combine phylogenetic reconstruction with optimality models taken from communication theory. One strength of evolutionary communication theory is that its principles are founded on existing mathematical methods, opening the door

to investigations of this kind.

Overall, evolutionary communication theory provides a framework for explanation of biological signalling behaviour. In the next section I argue that evolutionary biologists commonly employ *communicational thinking*, a set of assumptions that ground accepted explanatory practices.

7.4 Communicational thinking in biology

In this section I survey biological texts and highlight explanations given in terms of coordinated behaviour, divergent interests and physiological constraints. By treating sender and receiver strategies as evolutionary solutions to engineering problems, we help ourselves to a certain kind of adaptationist explanation. Such explanations are widespread in biology, and their salience goes unquestioned.

7.4.1 Trading effort and error

In chapter 6 I discuss a generalisation of two optimality approaches in communication theory. Rate-distortion theory and capacity-cost theory usually treat independent problems in communications engineering. The former theory deals with a trade-off between error and transmission rate, while the latter deals with a complementary trade-off between effort and transmission rate. Unifying them provides a general optimality model for communication, using transmission rate as an exchange rate between two kinds of cost. Studies of biological communication often pick up on aspects of this general framework.

Preece and Beekman (2014, p. 19), writing in the context of the waggle dance, speak

of "an inherent trade-off between effort and usefulness; communication is most worthwhile when the benefits of sending the signal exceed the cost of its production". This is the very same trade-off at the heart of rate-distortion theory (chapter 6). A few scientists explicitly consider the application of that theory in biological contexts (de Polavieja, 2002; Iglesias, 2016; Sims, 2016). But even those without knowledge of an existing mathematical framework would accept the logic described by Preece and Beekman.

The trade-off between effort and error emphasises the sender in a signalling interaction. But the receiver has a role too, as Endler makes clear:

> Receiving mechanisms probably evolve to increase the efficiency and reliability of information reception whereas signals probably evolve to increase the efficiency of communication and reliability of manipulation of the receiving individual to the benefit of the emitter.
>
> Endler (1993, p. 215)

These two selective forces come apart when interests diverge. Intuitively, common interest determines the extent to which these two selective forces act in tandem or conflict with each other.

Similarly, Marian Stamp Dawkins (1993, p. 251) discusses three selection pressures on signals that may sometimes push in different directions: "the degree of conflict or cooperation, the 'efficacy' of signals and how signalling costs are paid." She raises the difficulty of devising *general principles of signal design* due to the variety of selection pressures acting on communicative interactions. However, if I am right, there are at least four such principles. Dawkins raises a challenge for unifying frameworks such as evolutionary communication theory: they must "explain the diversity in size of animal signals and they must explain the diversity in form of signals" (Dawkins, 1993, p. 251). But evolutionary communication the-

ory says nothing about explaining diversity in size or form. It is a framework within which the properties of *particular* signals can be investigated. That being said, there is a sense in which the theory ought to apply 'globally'. Where biological signals are observed, it ought to uncover the mechanism(s) maintain them; where signals are *not* observed in social interactions (or where they quickly collapse upon being established), the theory ought to explain why. In the best case, the theory ought to generate predictions about which social interactions will be mediated by signalling. It is then a test of the theory whether signals do indeed appear in those situations.

7.4.2 Adapting to constraints

Romer (1993) considers long-range signalling and hearing in acoustic insects, particularly the role of constraints in explaining signalling strategies. Consider an example of a simple explanatory principle of this kind:

> Given that small size forces most insects to use higher frequencies, with the inherent problem of excess attenuation, we might predict that behavioural strategies (e.g. optimal broadcast positions), would be of significance in the evolution of acoustic signalling. Because scattering of sound within the vegetation seems to be the major cause of attenuation, insects should call from the top of the vegetation or close to it.
>
> Romer (1993, p. 180)

In support of this hypothesis, Romer cites observations of the bog katydid (*Metrioptera sphagnorum*) which calls from around 60cm up isolated spruce trees. Higher trees are available, but the diminishing returns of increased broadcast area versus climbing effort (along with other constraints such as increased exposure to predators) explain why that particular

height is optimal. This choice of height is a component of efficacy design. It is a means of countering noise inherent to signals pitched in the frequency range available to insects.

Romer (1993, p. 180) further notes features of insect signals that can be explained by reference to specific features of the environment: "Broad-band signals appear to be less vulnerable to amplitude fluctuations than are pure-tone signals [and] most bushcricket and grasshopper sound signals are broad-band rather than pure-tone." Insect signals are typically highly redundant, comprising continuously repeated phrases. Repetition is a rather unsophisticated (yet effective) means of counteracting environmental noise. Furthermore, certain signal features are unsuitable for indicating certain referents. Species identity, for example, is often a crucial component of signal variation. It would be inefficient to recruit a noise-prone aspect (such as tooth impact rate) to indicate the sender's species, except on small spatial scales (Romer, 1993, p. 180). All these considerations are the result of communicational thinking: receiver behaviour is assumed to be conditioned on signals; signals are adversely affected by interference; therefore, signal design can be explained by reference to the properties of the environment that constitute noise. Mechanical features of stridulation can be explained by reference to these communicational features (Montealegre-Z, 2009).

7.4.3 Efficacy design in bird calls

Maynard Smith and Harper (2003, §5.2) describe how the function of a call can explain its acoustic features:

> 'Flee' alarm calls, given to a cluster of individuals in immediate danger of attack, for example from a hawk, and causing all individuals rapidly to disperse, we would expect to be hard to locate, and not necessarily audible from a distance. 'Assembly' alarm calls cause hearers to assemble from a wide area, perhaps to

> mob a potential predator: we would expect them to be easy to locate, and audible at a distance.
>
> Maynard Smith and Harper (2003, p. 74)

The authors cite Marler (1955) as providing evidence for these hypotheses.

These considerations accord with principles about both joint function and divergent interests. The function of the Assembly call explains its form: wide-ranging, but easy to locate. The risks associated with eavesdropping explain the form of the Flee call too. Hawks use Flee calls as cues, leading to cryptic design: short-ranged, and hard to locate. Perhaps cryptic design ought to be considered a form of efficacy – or 'anti-efficacy' design, since they are in a sense diametrically opposed.

7.4.4 Social recognition and quality signalling

Social recognition is the use of cues and memory to assess conspecifics. Quality signalling is transmission of the same informational content, but via signals rather than cues. Sheehan and Bergman (2016, p. 2) treat social recognition and quality signalling as two distinct "assessment strategies", and investigate the "functional and evolutionary relationships" between them. A simple initial account was given by Rohwer (1982): in small groups, social recognition suffices; in large groups, memory constraints entail a need for quality signalling. Sheehan and Bergman (2016) expand on that earlier work and lay the groundwork for quantitative models. Their suggestions derive from the simple idea that social recognition and quality signalling are two means to the same end, so selection likely favours at most one of them. As a corollary, when the mechanism underpinning one conflicts with the mechanism

underpinning the other, selection ought to favour one over the other.[7]

For example, in birds, brain size is negatively correlated with pheomelanin plumage (Sheehan and Bergman, 2016, p. 9). While brain size can be used as a proxy for social recognition capabilities,[8] pheomelanin is known to play a role in signalling quality. We expect to see selection favour one of these features at the expense of the other, depending on social and environmental factors in the evolutionary history of each bird species. Just as with differing visual capabilities in surfperch, and adaptive radiation in leaf-nosed bats, variation in strategy investment in birds can be explained by reference to the goal of signalling (and cue reading) and inherent constraints on the strategies involved.

7.5 Communication theory in biology

In this section I survey biologists who explicitly refer to communication theory in explanations of signalling behaviour, showing that their approach is best interpreted by evolutionary communication theory.

7.5.1 Behavioural ecology

Among behavioural ecologists, the most concerted effort to employ mathematics in the study of communication has perhaps come from R. Haven Wiley (1983, 1994, 2013a, 2017). Over several decades, Wiley has argued for the direct application of the central model to animal

[7] On the other hand, if receivers recognise the cue and the signal with the same perceptual mechanisms, selection for one might entail improved performance with regard to the other. And if the information is extremely important, it might be worth maintaining both kinds of ability to ensure robustly reliable recognition.

[8] Two (significant) caveats: first, if brain size is a proxy for cue-reading then it is also likely a proxy for signal-reading; second, using brain size as a proxy for anything is often difficult to substantiate (Healy and Rowe, 2007).

communication. Wiley (1983) claims signals literally encode relevant aspects of the world such as the sender's suitability as a mate, strength, or willingness to fight. He also discusses game-theoretic considerations, such as the evolutionary stability of strategies in Prisoner's Dilemma games. Later, Wiley (1994) considers the trade-off between false positives and false negatives that shapes receiver strategies, drawing on results in signal detection theory. These investigations are closely related to rate-distortion theory (chapter 6), though Wiley does not appear to be aware of that approach. More generally, Wiley (2013a) justifies the use of concepts and models from communication theory in the study of animal signalling. Finally, Wiley (2017) describes twelve principles (or "predictions") about the evolution of communication in the presence of noise.

The particular mathematical approach Wiley favours is signal detection theory. That can be thought of as a receiver-focused subset of communication theory. However, this focus on receivers as individual decision makers backgrounds the significance of signalling as a joint enterprise. For example, Wiley treats divergence of interest as a source of noise. As we saw in chapter 6, that can be a useful perspective to take when answering questions about receiver behaviour. But in order to understand properties of the system as a whole, it should be treated as a joint enterprise that simultaneously optimises transmission efficiency and accurate response. Furthermore, instead of applying results from signal detection theory directly to biological situations, it would be better to develop a version of the theory that incorporates physiological constraints. For example, in contexts where receiver sensitivity derives from a pre-existing cue-reading adaptation, we can better understand the transition to signalling by figuring out which kinds of variation were likely to occur. Wiley (1994, p. 177) mentions external constraints on signal form such as the need to avoid alerting predators.

He is aware of the various forces that impact signalling behaviour, and his approach is an important step forward for understanding biological communication. It would benefit from incorporating ideas I have argued belong to evolutionary communication theory.

Another theorist who welcomes communication theory in biology is Jack Hailman. Hailman (2008) uses coding strategies as an organising principle. He categorises evolutionary and artificial signals with respect to the coding strategies they employ. For example, *discretised signals* are those that use discrete values to encode continuous referents (Hailman, 2008, Table 3.5 p.116). Artificial signals of this type include digital watches, digital speedometers and digital radio tuners. Hailman identifies several animal signals that fit this description too. For example, the South American electric fish *Sternopygus macrurus* produces electrical impulses using a muscle near the tail. Juveniles pulse at frequencies proportional to their size, which can be treated as a continuum. Adults, however, only emit pulses within a small range of frequencies, depending on their sex. The pulse continuum has effectively been discretised into two sets, indicating sex rather than length. Other examples appear in Hailman's table 2.1 (pp. 72-3), table 3.5 (116-7) and table 4.4 (160-1).

The success of Hailman's project rests on the contention that coding principles are independent of the modality of communication. Signals designed by humans and signals designed by evolution can be understood through a common functional analysis. Signals produced in the visual, auditory, chemical or other domain can all be analysed in the same way. Hailman accepts that specific measures of surprisal, entropy and transmission rate are not as important as the conceptual organisation provided by the engineering approach. What is most important is the common form taken by explanations of signalling strategies. At its broadest, the optimism Hailman endorses pushes towards a generalised view of commu-

nication theory. Because communication is a phenomenon distinguished by its functional shape – codesigned senders and receivers whose joint functioning is mediated by relatively low-energy states – the same design principles apply to every instance of communication, whether natural or artificial.

Hailman's approach would be improved by acknowledging the principles of evolutionary communication theory. Coding schemes in nature are not freely chosen, but are often the result of physiological constraints. For example, we might imagine that the discretised code of the electric fish originated as a side-effect of sexual dimorphism: females, being smaller, tended to produce pulses whose average frequency was noticeably different from those of males. Of all the cues that might have been enlisted by selection to become signals, this was the one most amenable to tinkering. In this hypothetical scenario, the 'code' is largely a by-product, which was only refined late in selection history to improve detection of sex differences. Furthermore, in a different version of the hypothetical scenario, suppose divergence of interest between competing males led some to adopt a strategy of pulsing at female frequencies.[9] This allows them to sneak through the territories of other males without risk of attack. Widespread adoption of this strategy would lead to counter-selection for better discrimination ability. The 'code' is then not purely a consequence of cooperative selection, but has been refined by selection acting on competing senders and receivers.

In sum, while Hailman adopts an idealised perspective on animal signalling codes, we can pursue more realistic analyses by adopting the perspective of evolutionary communication theory.

[9] Male cuttlefish are known to adopt a strategy of this kind (Hanlon et al., 2005).

7.5.2 Molecular signalling

In the microscopic domain, mathematical work is burgeoning in the relatively new field of molecular communication theory. Two related approaches fall under this heading: understanding microbial signalling behaviour with respect to engineering concepts, and developing microscopic communication systems. Adam Noel and colleagues take the first approach, providing a simple justification for interpreting biological communication through the lens of engineering:

> Signals are being regularly transmitted within and between individual cells and microorganisms. These signals may not be sending packets of data in the conventional communication sense, but nevertheless they enable conventional communication applications such as sensing, coordination, and control. Thus, we can adapt conventional communication engineering theory and techniques to study these signaling mechanisms and understand how they work.
>
> Noel et al. (2017, p. 1)

Biological entities do not instantiate the central model – they do not reconstruct symbol strings – but they do use signals to achieve the same goals that humans developed the central model for in the first place. The concepts of communication theory apply in both domains because the applications are the same, even if the mechanical details are not. The argument suggests a general motto in line with the considerations we have developed throughout this thesis: *it is the function, not the mechanism, of communicative behaviour that justifies using concepts, methods and tools from communication theory in biology*.

Studies of molecular signalling would benefit from adopting the perspective of evolutionary communication theory. Explicit consideration of divergence of interest is lacking within communication-theoretic models, because it has not previously been clear how to generalise

communication theory to non-cooperative contexts. Such ideas exist in game-theoretic approaches to bacterial communication (Diggle et al., 2007a). Furthermore, taking account of constraints on signalling and cue-reading could illuminate the dynamics of microbial social interactions. Bacteria are likely far more constrained in their behaviour because of limited (perhaps absent) ability for associative learning. But their relatively faster rates of evolution – plus the possibility of horizontal gene transfer – can lead to interesting social dynamics. Overall, studies of microbial communication would do well to consider the evolutionary framework described in this chapter.

7.5.3 Genetics and evolution

Bergstrom and Rosvall (2011) argue for interpreting the genetic code as a solution to traditional communication-theoretic problems. An inheritance system must reliably transmit instructions across generations which can be read in development. The specific problem facing the transmission part of this two-step process is to efficiently 'package' those instructions, avoiding costly mistakes induced by noise or misreading. The authors argue that the genetic code shows features designed specifically to counteract or mitigate such untoward events (Bergstrom and Rosvall, 2011, pp. 173–4). This is channel coding, as opposed to source coding (cf section 5.5 and appendix A). They also reject the Shannon/semantic distinction, framing their own "transmission" interpretation in contrast to both those perspectives (cf section 3.5). Although in the present work we are not concerned with genetics, it is salient to mention that many of the points I am trying to support have similar justification in that context.

Once again, this work would benefit from explicitly considering principles of evolution-

ary communication theory, particularly failures of common interest. For example, outlaw genes are those that reproduce at the expense of the fitness of genome of which they are a part (Alexander and Borgia, 1978). By exploiting the properties of mitosis, outlaws are able to get themselves over-represented in the gametes that will become the next generation. These genes do not promote the fitness of the organism. Indeed, they hamper the fidelity of the inheritance system conceived as a process designed for information transmission. The question whether outlaw genes can be fruitfully considered 'noise' – and, if so, how the system ought to be designed to combat it – is one that could be asked from the standpoint of evolutionary communication theory.

In separate work, Bergstrom and colleagues have pushed for biological interpretations of definitions and equations originally presented in a communication theory context. Both Bergstrom and Lachmann (2004) and Donaldson-Matasci et al. (2010) connect biological fitness (interpreted as growth rate of a lineage) with the transmission rate of a cue channel. Inspired by the definition of transmission rate given by Kelly (1956), the authors combine utilities and probabilities to define the increase in fitness that results from conditioning a behavioural strategy on an environmental cue. Kelly's exposition concerns a gambler's winnings. By casting that work in biological terms, Bergstrom and colleagues offer an explicit justification for applying mathematical measures to biological cues. To my knowledge they do not extend the interpretation to signals, nor do they investigate rate-distortion mathematics that apply to communication rather than cue-reading. Nonetheless, their mathematical treatment prompts a literal interpretation of engineering definitions in biology.

Overall, a minority of biologists embrace the prospects for communication theory in biology. My hope is that evolutionary communication theory provides positive branding for

their cause, assimilating their ideas around a single label and highlighting its relationship to already well-established modelling frameworks like game theory.

7.6 Objections

7.6.1 Complexity: It is too difficult to calculate and/or measure the variables mentioned in communicative trade-offs

Objection. The variables in proposed models of evolutionary communication include functional success, noise, metabolic expenditure, physiological constraints and divergence of interests. Even if we could adequately put these together into a model, it would not be possible to experimentally test for them in a given situation. Therefore, the proposed framework is intractable, and cannot inform biological practice.

Response. Let us first clarify the objection. In particular, notice how it differs from Owren, Rendall and Ryan's (ORR's) objection to communication theory in behavioural ecology. ORR complained that information was not an appropriate concept for animal communication. Their reasons included the effect of constraints and divergence of interests in signalling systems. Once it has been shown that these factors can be accounted for within an expanded, evolutionary communication theory, ORR's initial challenge has been met. The current objection is different. It says that the modelling practices implied by evolutionary communication theory are too difficult to implement.

As a second clarificatory point, note that the objection is not to the *accuracy* of models in the proposed framework. That a model is difficult to implement cannot be a mark against a theory's validity. If evolutionary communication theory did not adequately describe the

trade-offs that explain signalling behaviour, that would be one thing. But the complaint is about complexity rather than accuracy.

Now to respond to the objection proper. Far more complex modelling frameworks than the one proposed here have been implemented in biological research. Evolutionary communication theory considers four fundamental aspects of signalling behaviour (coordinative function, environmental noise, divergent interests, physiological constraints). In comparison to some evolutionary models, this is positively sparse. I agree that such features would be hard to measure. But their initial use would be like that of the terms in Hamilton's rule: employed in idealised models and as part of an organising framework, only later measured or estimated when empirical methods improve (Bourke, 2014; Gorrell et al., 2010).

Furthermore, I do *not* suggest that this is a novel modelling framework that can explain old phenomena in new ways. Rather, I suggest that *existing patterns of explanation – sometimes well-established and widely accepted patterns – should be conceptualised within a framework such as that which I have described.* I have also given the framework a name, partly to highlight mathematical links with both communication theory and evolutionary game theory, but also to unify disparate research programs working with the same explanatory tools. I believe that interpretive issues around the term 'information' would largely disappear were this framework to be adopted.

Some further remarks are in order. Notice that this is not a question of choosing the simpler theory. The objection has nothing to do with Occam's Razor. For that to be the case, there would have to be competing theories that purport to explain the same phenomena. The accuracy of the theories would have to be reasonably similar, such that simplicity could act as a tie-breaker when scientists choose between them. That is not what is going on here. It

is not as though we have competing theories of signalling behaviour, and some alternative to evolutionary communication theory is simpler.

Rejoinder. Influence-based definitions of communication such as those promoted by Dawkins and Krebs (1978) and Owren et al. (2010) are alternatives to evolutionary communication theory. Suppose we could demonstrate that the influence-based definition entails a simpler theory of animal signalling. Then, by Occam's Razor we would have reason to reject evolutionary communication theory.

Response to rejoinder. If this were indeed the structure of the current dialectic, it might well be worth considering issues of simplicity. But that is not what the debate looks like. Influence-based definitions of signalling are not competitors to evolutionary communication theory. In particular, they do not posit different explanations of signalling behaviour. Rather, they place emphasis in different places. For one thing, evolutionary communication theory purports to describe signalling in a much wider range of taxa than the animal kingdom. For another, it seeks to distinguish cases based on their similarity to paradigmatic signals (chapter 1). The issues raised by ORR motivate an influence-based definition by sidelining cooperative signalling. But it can hardly be denied there are cooperative signals in nature.

Furthermore, as I argued in section 7.2 the considerations that prompted the influence definition ought to be incorporated into evolutionary communication theory. That is one major argument of this thesis: that insofar as ORR have raised relevant concerns about an implicit cooperative bias of animal signalling researchers, we ought to develop a theory of communication that does not take common interest for granted. I propose accepting the core motivation for ORR's analysis: that many (perhaps most) animal signals lie on the influence end of the common interest dimension in the signalling hypercube. The truth of

this statement does not threaten evolutionary communication theory.

7.6.2 Novelty: Evolutionary communication theory is not novel

Objection. By giving a name to an existing family of practices, I have done no real analytical work. Biological studies of communication got along fine without philosophers interfering. Indeed, by producing so many examples of biologists whose explanations I take to be aligned with the ideals of evolutionary communication theory, I have exposed that framework as nothing more than existing theory re-described. There is nothing novel in the approach I advocate.

Response. This objection would hit home were it not for the fact that biologists continue to argue about the relevance of information theory for their discipline. It is not yet appreciated what explanatory role the concept of information plays. By collecting extant theoretical approaches under a common framework, we put paid to claims that information is an idle metaphor. Moreover, we should not dismiss the power of names. Sometimes a small change in perspective, prompted by an apposite label, clears away years of theoretical impasse. In the present case my label completes the square depicted in table 7.1. It suggests we have no more reason to be wary of the concept 'information' than we do of concepts borrowed from game theory.

7.7 Conclusion

Engineers make rational decisions about the communication systems they construct. Signal form depends on factors including the function to be performed, the channel through

which the signal is to be sent, and constraints such as power availability. The same design principles guide the evolutionary emergence of biological signals. The most significant disanalogy between the mathematical theory and its evolutionary counterpart is the prevalence of imperfect common interest in biological interactions. However, even that feature can plausibly be incorporated into mathematical descriptions of signalling behaviour. The result is an illuminating conceptualisation of a well-established family of practices: evolutionary communication theory.

Conclusion

A bird sits on a telegraph wire and sings. What explains its song? For it to be worthwhile, the song must have an expected effect that brings an evolutionary benefit outweighing metabolic expenditure. This trade-off is tempered by environmental noise that limits the song's efficacy, divergent interests that incentivise its audience to act uncooperatively, and physiological constraints that restrict the song's viable forms. The theoretical approach that licenses such explanations is evolutionary communication theory. It is related to communication theory in just the same way as evolutionary game theory is related to game theory. Mathematical concepts and results first discovered in engineering apply to biological contexts.

In reaching this point, we made heavy use of mapping relations, a theoretical entity posited by sender-receiver teleosemantics. Mapping relations are not correlations or natural correspondences. They are functional specifications of relational properties that signals ought to bear. 'Ought' is to be understood in terms of proper function. Signals are special because they mediate two or more proper functions, and as such serve a joint proper function. Paradigmatic signals cause receiver behaviour, but the energy with which the behaviour is performed is supplied by the receiver rather than the signal. It is this feature that warrants an informational analysis. Mapping relations are the fundamental form of semantic content. Far from being distinct concepts, information and content pick out the same phenomena and play the same explanatory role. Ignorance of their identity owes much to misunderstandings of communication theory. The special features of the central model justify claims about the irrelevance of meaning for informational measurements. Such claims cannot be generalised to biological contexts.

This thesis argued for the unification of theories from three disciplines: communication theory from engineering, evolutionary communication theory from biology, and teleosemantics from philosophy. When put together, all three shed light on the concept of information. More importantly, they offer insight into the fundamental nature of meaning, and its place in our material world.

Appendix A

The central model of communication theory

Definitions

The central model of communication theory is a causal model with six nodes (figure A.1). The **source** generates a string w that is interpreted as a message to be transmitted to the **target**. The **encoder** converts source message w into codestring x. Upon being transmitted through the **channel**, codestring x may be altered by **noise**. The received string, which may or may not be identical to x, is labelled y. The **decoder** converts y into z.

According to Shannon (1948b, p. 379), "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." This is equivalent to making z match w. In a sense, the choice of z is fully under the decoder's control. Because it is designed in tandem with the encoder, the decoder need only have access to the transmitted string x in order to reconstruct w to a desired degree of accuracy. First, consider how *predictable* x is. If the transmitted string were always the same, the decoder would always produce the same z. The task would be trivial. As x gets

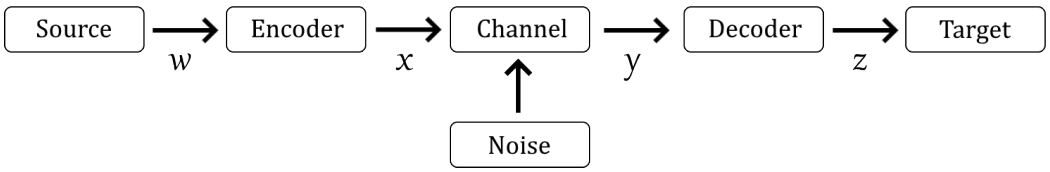

Figure A.1: The central model of communication theory.

less and less predictable, however, the decoder will be less accurate when it simply guesses. A useful measure will be *how unpredictable* the transmitted message is.

Shannon's definition of information is a way to measure unpredictability of a message. That an event x has probability $p(x)$ is familiar enough. We can further associate a measure of unpredictability to the event, $\log \frac{1}{p(x)}$. Formally, we call this quantity **surprisal**. Events with *lower* probability are *more* surprising, which explains why the term $\frac{1}{p(x)}$ appears in the definition. But why is a logarithmic function applied?

Consider an example. For two independent outcomes each with probability $\frac{1}{2}$, the probability of both occurring is $\frac{1}{4}$. If the measure of surprisal were simply $\frac{1}{p(x)}$, then the surprisal of either event alone would be 2 and the surprisal of both together would be 4; the surprisal of three events would be 8, of four 16, and so on. Surprisals would multiply as probabilities multiply. This might be fine for some uses, but there is a special role we want our definition of surprisal to play. We want it to tell us how many symbols would be required to encode the outcomes of these events. For two events, each with two possible outcomes, there are in total four different possible outcomes, but we only need two binary symbols to distinguish all four of these. As figure A.2 demonstrates, the number of symbols required to distinguish between outcomes grows only arithmetically while the number of outcomes grows geometrically. In other words, surprisals should *add* as probabilities multiply. Logarithms relate things that grow geometrically with things that grow arithmetically. So, when figuring out how many symbols are required to encode a given sequence of outcomes, logarithms are the right tool for the job.

When dealing with a collection of mutually exclusive and jointly exhaustive events – an *event space* – we wish to know how surprised we should expect to be. The answer is

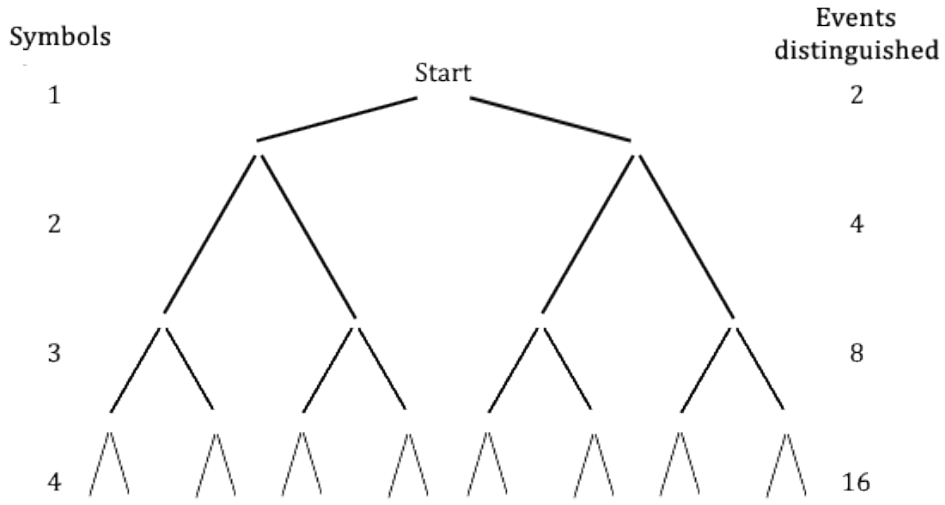

Figure A.2: A pictorial explanation for why the units of information are logarithmic. Defining a *binary symbol* as one which distinguishes two equiprobable events, a sequence of N binary symbols distinguishes between 2^N equiprobable events. The mathematical operation that retrieves the number N from the number 2^N is the logarithm to base 2: $\log_2 2^N = N$. In general, $\log_a a^N = N$.

a weighted sum of the surprisals: with probability $p(x_1)$ we will be surprised by $\log \frac{1}{p(x_1)}$, with probability $p(x_2)$ we will be surprised by $\log \frac{1}{p(x_2)}$, and so on. Summing the surprisals weighted by the probabilities yields the expected surprisal, or **entropy**, of the event space. It can be thought of as the uncertainty regarding the set of events in question. Entropy is greater when there are more events and/or events have similar probabilities, and it is lower when there are fewer events and/or a small proportion of those events have significantly greater probabilities than the others. Its formal definition is $H(X) = \Sigma p(x) \log \frac{1}{p(x)}$.

Because the received string y is causally connected to the transmitted string x, uncertainty about x will be reduced when y occurs. A measure of the remaining uncertainty about x once y has been observed is the **conditional entropy**: $H(X|Y) = \Sigma p(x,y) \log \frac{p(y)}{p(x,y)}$. A reasonable definition of how much uncertainty has been reduced by the process of communication is the prior uncertainty minus the residual uncertainty after transmission. This is **transmission rate**: $R = H(X) - H(X|Y)$.

Transmission rate is determined partly by conditions in the channel, and partly by the choice of code. For this reason, it is helpful to define a value that is the maximum possible transmission rate across alternative codes. This is **channel capacity**: $C = \max_{p(x)} R$. The term $\max_{p(x)} R$ means the maximum value of R that can be achieved by changing the distribution of transmitted symbols $p(x)$. If some symbols are used much more commonly than others, the code could be made more efficient by altering it such that all symbols are used equiprobably. An example will help make this clear.

Consider a source that produces four symbols w_1, w_2, w_3 and w_4.[1] This can be inter-

[1]The example is taken from Martínez (2020), who uses it to illustrate principles of source coding (see below).

preted as the possible outcomes of a probabilistic event such as a horse race. Suppose the probabilities of the outcomes are not equiprobable, so:

$$p(w_1) = \frac{1}{2}$$
$$p(w_2) = \frac{1}{4}$$
$$p(w_3) = \frac{1}{8}$$
$$p(w_4) = \frac{1}{8}$$

Suppose we have a noiseless channel that supports sending signals composed of two symbols, 0 and 1. There are various ways we can assign each outcome to a sequence of symbols. Perhaps the most obvious is to use a unique combination of two symbols, like so:

$$w_1 \rightarrow 00$$
$$w_2 \rightarrow 01$$
$$w_3 \rightarrow 10$$
$$w_4 \rightarrow 11$$

Whatever virtues this code may have, it has one significant drawback: each symbol is not equiprobable. Because outcomes w_1 and w_2 are more probable than w_3 and w_4, the receiver already expects to observe a 0 in the first position. And having received a first 0, they would more often than not observe a second 0 too – indicating outcome w_1, the most probable. As a result, the entropy of X is not as large as it could be, and neither is the transmission rate:

$$
\begin{aligned}
p(0) &= \frac{\frac{3}{4} + \left(\frac{3}{4} \times \frac{2}{3}\right) + \left(\frac{1}{4} \times \frac{1}{2}\right)}{2} \\
&= \frac{11}{16} \\
p(1) &= 1 - p(0) = \frac{5}{16}
\end{aligned}
$$

$$
\begin{aligned}
H(X) &= \frac{11}{16} \log \frac{16}{11} + \frac{5}{16} \log \frac{16}{5} \\
&\approx 0.8960
\end{aligned}
$$

$$
R = H(X) - H(X|Y) = 0.8960 - 0 = 0.8960 \text{ bit/symbol}
$$

There is a code, however, that uses *fewer than 2 symbols per signal*, on average:

$$
\begin{aligned}
w_1 &\to 0 \\
w_2 &\to 10 \\
w_3 &\to 110 \\
w_4 &\to 111
\end{aligned}
$$

Now the probabilities of transmitted symbols has been 'equalised': the first position has equal chance of being 0 or 1; supposing the first symbols is a 1, the second position is also 50/50; and supposing the second symbol is a 1 the third position is also equiprobable. The transmission rate is therefore:

$$p(0) = \frac{1}{2}, \quad p(1) = \frac{1}{2}$$
$$H(X) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1$$
$$R = H(X) - H(X|Y) = 1 - 0 = 1 \text{ bit/symbol}$$

For a noiseless channel, equalising the probabilities of transmitted symbols maximises transmission rate.

In the first code, the average length of a signal is 2 symbols (because all signals have this length). In the second code, on the other hand, the average length of a signal is 1.75 symbols. By equalising probabilities, we reduce the average length of transmission (thus saving on signalling resources) without jeopardising the integrity of the message.

The definitions introduced in this section are collected in table A.1 along with their usual interpretations in the central model.

Table A.1: Definitions used in communication theory.

Term	**Mathematical interpretation**	**Intuitive interpretation**
W	Source string. A random variable.	Set of possible source messages
w	A specific source string. A value of W.	Message to be sent
X	Encoded string. A random variable.	Set of possible transmitted codewords

Continued on next page

Table A.1 – continued from previous page

Term	Mathematical interpretation	Intuitive interpretation
x	A specific encoded string. A value of X.	The transmitted codeword
Y	Received string. A random variable.	Set of possible received codewords.
y	A specific string received. A value of Y.	Signal received (may differ from x due to noise in the channel)
Z	Decoded string. A random variable.	Set of possible decoded messages.
z	A specific decoded string. A value of Z.	Reconstructed message (may differ from w due to noise in the channel or coding error)
$\log \frac{1}{p(x)}$	Surprisal of event x	Uncertainty reduced upon learning x
$H(X) = \Sigma p(x) \log \frac{1}{p(x)}$	Entropy of transmitted string.	How uncertain we are about the transmitted string.

Continued on next page

Table A.1 – continued from previous page

Term	Mathematical interpretation	Intuitive interpretation
$H(X\|Y) = \Sigma p(x,y) \log \frac{p(y)}{p(x,y)}$	Conditional entropy of transmitted string given received.	How uncertain we are about the transmitted string *after* having observed the received string.
$R = H(X) - H(X\|Y)$	Transmission rate of the channel.	Initial uncertainty about the transmitted string $H(X)$ minus the residual uncertainty $H(X\|Y)$.
$C = \max_{p(x)} R$	Capacity of channel.	Maximum transmission rate.

Source coding and channel coding

Source coding compresses strings. It removes redundancies – patterns – in a string in order to record it as efficiently as possible. For example, suppose your string is five hundred 1s followed by five hundred 0s. It would be a waste of space to transmit all one thousand symbols if there was a shorter way of transmitting the description *five hundred 1s followed by five hundred 0s*. Source coding offers solutions to the problem of efficient use of signalling resources.

Channel coding prepares strings for transmission. It judiciously *adds* redundancy to a

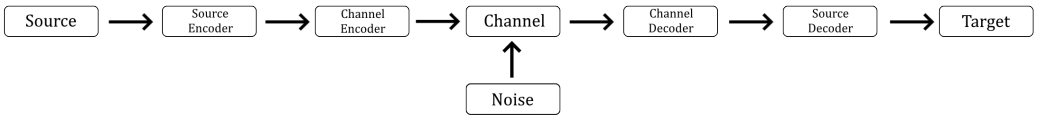

Figure A.3: The extended central model. The source encoder removes redundancies in source string, while the channel encoder judiciously *adds* redundancy to counter noise in the channel.

string, protecting the message from noise. Chosen well, channel coding ensures that even if the codestring is distorted by noise, the original can still be recovered. For example, suppose you send the same signal five times. Interference might affect each of those five copies in different ways, but by taking the 'average' over all five received signals, the decoder might well be able to recover the original string.

In practical contexts, source coding and channel coding are combined. Sending *five hundred 1s followed by five hundred 0s* five times might take less time and power then sending the entire sequence once. It will certainly be more efficient than sending the entire sequence five times. The central model can be extended by distinguishing source and channel coding (figure A.3).

Consonant with the two varieties of coding are two definitions of rate. **Source rate** (also called *information rate*) is the average entropy per symbol produced by the source. In our terminology, this is $H(W)$. For example, when linguists speak of the 'entropy' of a language such as English, they refer to this value. The population of English speakers is modelled as a source. The unpredictability of units of language at a particular level – phonemes, morphemes, words, phrases – defines the source rate, which may be estimated by studying corpora (Bentz, 2018). **Channel rate** (which we have called *transmission rate*) is the average reduction in uncertainty per symbol transmitted through the channel. As above,

this is $H(X) - H(X|Y)$.

Theorems

Two fundamental theorems establish the limits of source coding and channel coding. Each has two versions, one in which error is prohibited and one that considers a predefined level of tolerable error.

The **source coding theorem** provides an answer to the following question. Given a sequence of events drawn independently from the same source, how many symbols are required to record those events with negligible error? We might imagine a source producing symbols at a certain rate, generating a string which we must compress by encoding. Encoding generates a new string, which can be used to recover the original (assuming no noise). We want to use as few binary symbols as possible in our encoded string: what is the minimum we can achieve? The answer is given in terms of the entropy of the event space and the number of events in question:

> Let the event space have entropy H, and the number of events to be encoded N, then as N tends to infinity the number of binary symbols required is NH.

Proofs can be found in MacKay (2003, §4.5), Cover and Thomas (2006, §5.4), and Shannon (1948b, §9).

The **source coding theorem *with error*** provides an answer to the following question. Given a sequence of events drawn independently from the same event space, how many symbols are required to record those events with tolerable error? The problem is equivalent to that given above, except we have some margin for error in the reconstructed string. Here error is measured as a function of the difference between each symbol of the original string

and its reconstructed counterpart.

Shannon (1959) showed that there will often exist a function, $R(d)$, that is the 'equivalent rate' of a source when errors are allowed. Here d is a measure of the cost of reconstructing the wrong symbol, and $R(d)$ is a function of it. The same source will therefore have different equivalent rates depending on how much error is tolerated. At one extreme, a source for which every symbol can be in error has equivalent rate zero. This is because there is no restriction on the process by which the sequence is decoded. At the other extreme, a source for which every symbol must be reconstructed accurately has equivalent rate equal to the original source rate. Then the problem reduces to the original source coding theorem. These issues are treated by rate-distortion theory (see chapter 6).

The **channel coding theorem** (often called the noisy channel coding theorem) answers the following question. What is the maximum number of symbols that can be reconstructed, with negligible error, per use of a channel with noise? Shannon (1948b, §13) showed that the answer is C, the channel capacity. Again, this theorem links a probabilistic measure with a quantity of symbols. Lore has it that before Shannon's foundational text, engineers widely assumed that noisy channels would lead to a non-negligible probability of error no matter how low the rate of transmission. Shannon's proof ushered in an era of optimism. He demonstrated that there were in principle codes that could shield messages from noise, to a certain measurable level.

Finally, the **channel coding theorem *with error*** answers the following question. What is the maximum achievable rate R given an allowable transmission error d? As with the corresponding source coding theorem, the answer derives from a function $R(d)$ where d is a measure of tolerable error. MacKay (2003, §10.3) proves that when d is the probability of

bit error, the maximum rate is:

$$R(d) = \frac{C}{1 - H_2(d)}$$

where H_2 is the binary entropy function (MacKay, 2003, p. 2):

$$H_2(x) \equiv x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$$

Because we are now speaking of rate in a *channel* context, rate can increase beyond the capacity of the channel. This raises a conceptual question: what does it mean for transmission rate to exceed capacity? Naively, such a thing is impossible: capacity is *defined* as the maximum possible rate (table A.1). Contradiction is avoided by defining R as the maximum tolerable ratio of symbols we attempt to reconstruct to symbols being sent through the channel. The original definition, $H(X) - H(X|Y)$, assumes that symbols must be reconstructed perfectly. When error can be tolerated, transmission rate – the number of symbols we attempt to reconstruct per channel use – can increase beyond this. Defining rate as the ratio $R \equiv \frac{\text{reconstructed symbols}}{\text{transmitted symbols}}$, the statement $R = H(X) - H(X|Y)$ becomes a theorem rather than a definition. It holds only in the special case of zero tolerable error.

Appendix B

Signalling games

A **signalling game** is a quadruple $G = (P(W),\ Q_s,\ Q_r,\ m)$ where $P(W)$ is a normalised vector of length l, interpreted as the unconditional probability distribution over states W; Q_s is an $l \times n$ matrix, interpreted as the sender's **payoff matrix**, Q_r is a $l \times n$ matrix, interpreted as the receiver's payoff matrix, and m is the number of signals. The rows of each payoff matrix correspond to a state, and each column corresponds to an act. Each cell denotes the payoff of each unique state-act combination.

Games are schemas within which the properties of different **strategies** can be evaluated. In signalling games, the strategies of senders and receivers are represented by **strategy matrices** $P(S|W)$ (with dimension $l \times m$) and $P(A|S)$ (with dimension $m \times n$) respectively.

Games model strategic interactions. Games are not strictly causal models, though causal models can represent strategic interactions too. Although in the main text I use the term **sender-receiver model** to refer to causal models of signalling, for our purposes there is no difference between signalling games, sender-receiver games and sender-receiver models; I note the difference here for the sake of correctness. Finally, the **sender-receiver framework** is the collection of sender-receiver games and causal models, plus hypotheses, explanations, observations and predictions derived from them.

Populations and individuals

Signalling games may represent populations of agents or just one agent per role (sender and receiver). Either way, the mathematical objects used to represent payoffs and strategies are the same. A strategy matrix defines the probabilities of acting in certain ways under certain conditions. It may be interpreted as a single individual acting stochastically, or as a population of agents all individually deterministic but with different strategies represented at differing proportions. In the population case, probabilities in the strategy matrix denote what behaviour one should expect to see when choosing an agent from the population at random.

Typically, strategy matrices change over time in accordance with a **dynamic equation** which increases the frequency of a strategy in proportion to the payoff it brings. A dynamic equation is simply a rule that states what happens at a future time step given what happened in the past. Perhaps the most familiar dynamic equation is the *replicator equation*:

$$x_i' = x_i . \frac{f_i}{\overline{f}}$$

Where x_i' is the proportion of strategy i at the next timestep, x_i is its proportion at the current timestep, f_i is its current payoff, and $\overline{f}$ is the average current payoff in the population. The equation specifies how many copies of a strategy appear in the next population, given that strategy's relative success in the present environment.

Success f_i is usually calculated as a simple sum of the payoff matrix components. The game is played for one or more **rounds** per timestep. During each round an interaction occurs between sender and receiver. Depending on the outcome of the interaction, payoffs are assigned to each sender/receiver strategy. When the total number of rounds has been played,

the timestep increases by one, and the replicator equation determines the new distributions of strategies.

In what follows, I will continue to speak as though signalling games represent one agent per role.

Game dynamics

One round of a signalling game proceeds as follows. A state occurs with probability $p(w_i)$, a signal is chosen with probability determined by the state and the sender strategy distribution, and an act is performed with probability determined by the signal and the receiver strategy distribution. Payoff matrices determine payoffs for combinations of state and act.

The sender's strategy matrix $P(S|W)$ determines how signals are chosen on the basis of observed states. Rows correspond to states while columns correspond to signals. The matrix therefore has the dimensions $l \times m$. Cell p^s_{ij} is the probability that the sender will send signal s_j upon observing state w_i. The receiver's strategy matrix $P(A|S)$ determines how acts are chosen on the basis of observed signals. Rows correspond to signals while columns correspond to acts. The matrix therefore has dimensions $m \times n$. Cell p^r_{jk} is the probability that the receiver will perform act a_k upon observing signal s_j.

Strategies change over time. At a timestep, $P(S|W)$ and $P(A|S)$ represent the probabilities of sender and receiver producing signals and performing acts respectively. We can calculate the conditional matrix:

$$p(a_k|w_i) = \Sigma_j p(a_k|s_j)p(s_j|w_i)$$

Now $p(a_k|w_i)$ is the probability, for a population at a time, that act a_k will be performed

given state w_i occurred. We already know the unconditional probabilities of states $P(W)$ because they are specified as part of the game. We can therefore derive the joint distribution $P(A,W)$:

$$P(A,W) = P(A|W)P(W)$$

and we can derive the unconditional probabilities of acts $P(A)$ by marginalising: $p(a_k) = \Sigma_i p(a_k, w_i)$.

All this together allows us to calculate both the entropy of states and the conditional entropy of states given acts:

$$H(W) = \Sigma p(w_i) \log \frac{1}{p(w_i)}$$
$$H(W|A) = \Sigma_{i,k} p(w_i, a_k) \log \frac{p(a_k)}{p(w_i, a_k)}$$

There is therefore a well-defined notion of **transmission rate** at a timestep of a sender-receiver game:

$$R = H(W) - H(W|A) = I(W;A)$$

Perhaps the least controversial way to interpret this term is as *what an external observer could learn about the act to be performed given that they know the state, or vice versa*. No assumption of cooperation on the part of the agents is required for this definition to hold, or for transmission rate to be greater than zero. Payoffs might be such that sender and receiver have fiercely competing interests; nonetheless, at some point during the game a statistical connection may occur between states and acts. Such a connection could be a consequence

of one party exploiting the other. If interests diverge far enough we might eventually see the exploited party change their behaviour, and the connection will be lost. Wagner (2012) gives a model of a zero-sum game in which cycles of exploitation continue indefinitely.

Notice the subtle but important difference between transmission rate in signalling games and in the central model (appendix A). There, the variables over which transmission rate was calculated were X (the encoded string that is about to be transmitted) and Y (the received string that is about to be decoded). The equivalent transmission rate in a signalling game (assuming no noise) is just $H(S)$, the entropy of signals. This places an upper bound on $I(W;A)$: you cannot transmit more information through a channel than that channel's intrinsic variation.

A line of future research is in determining whether anything general can be said about $I(W;A)$ solely on the basis of payoff matrices and the unconditional probabilities of states. Intuitively, $I(W;A)$ should be lower when interests diverge and higher when they coincide. Martínez and Godfrey-Smith (2016) present some results along these lines. It seems inevitable there should be analytical results, extensions to the theorems outlined in appendix A, that describe the limits on communication between conflictual agents.

A worked example

Consider the cooperative 2x2 state-act game:

$$\underbrace{P(W) = \left(\frac{1}{2}, \frac{1}{2}\right)}_{\text{States equiprobable}} \quad \underbrace{Q_s = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad Q_r = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\text{Perfect common interest}} \quad \underbrace{m = 2}_{\text{2 signals available}} \tag{B.1}$$

Suppose we begin with a randomised strategy pair:

$$\underbrace{P(S|W) = \begin{pmatrix} \frac{1}{6} & \frac{5}{6} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}}_{\text{Sender strategy}} \quad \underbrace{P(A|S) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{12} & \frac{11}{12} \end{pmatrix}}_{\text{Receiver strategy}}$$

Strategies like this are called **mixed strategies** because they can be thought of as linear combinations of **pure strategies**. For example, the sender matrix can be written:

$$\underbrace{P(S|W) = \begin{pmatrix} \frac{1}{6} & \frac{5}{6} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{2}{12} & \frac{10}{12} \\ \frac{9}{12} & \frac{3}{12} \end{pmatrix}}_{\text{Mixed strategy}}$$

$$= \underbrace{\frac{1}{12}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \times \frac{8}{12}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \times \frac{1}{12}\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \times \frac{2}{12}\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}}_{\text{Linear combination of pure strategies}}$$

In each round, one of these strategies is selected according to its probability ($\frac{1}{12}$, $\frac{8}{12}$, $\frac{1}{12}$, $\frac{2}{12}$). If Sender is successful, that strategy's payoff for this timestep is increased by 1. Suppose there are 1000 rounds per timestep and the respective strategies enjoy the following payoffs:

$P(S\|W)_1$	$P(S\|W)_2$	$P(S\|W)_3$	$P(S\|W)_4$
$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$
441	127	148	200

84 turns resulted in no payoff. (These numbers are not necessarily plausible given the strat-

egy pair and payoffs; I am using them for illustration only.) Now, the replicator equation takes these payoffs and the proportions of each strategy and generates new proportions. Take for example $P(S|W)_1$:

$$x_1 = \frac{1}{12} = 0.08333... \text{ (starting proportion of } P(S|W)_1)$$

$$f_1 = 441 \text{ (payoff for } P(S|W)_1)$$

$$\overline{f} = \frac{441+127+148+200}{4} = 229 \text{ (average payoff across strategies)}$$

$$x_1' = x_1.\frac{f_1}{\overline{f}} = \frac{1}{12}.\frac{441}{229} = 0.1605... \text{ (new proportion of } P(S|W)_1)$$

Strategy $P(S|W)_1$ has almost doubled in frequency.

Converting a payoff matrix into a distortion matrix

Martínez (2019, p. 3) gives the following formula to convert payoffs into distortion measures. Consider a payoff matrix Q. Let $q_{\max} = \max_{ik} q_{ik}$ and $q_{\min} = \min_{ik} q_{ik}$. Then the distortion matrix corresponding to Q is defined by:

$$d_{ik} = \frac{q_{\max} - q_{ik}}{q_{\max} - q_{\min}}$$

For example (Martínez, 2019, Tables 1&2, p.4):

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad q_{\max} = 1, \quad q_{\min} = 0, \quad D = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

A more complicated example (Martínez, 2019, Table 3, p.5):

$$Q_s = \begin{pmatrix} 2 & 0 & 10 \\ 0 & 2 & 10 \\ 0 & 10 & 0 \end{pmatrix}, \quad q_{\max} = 10, \quad q_{\min} = 0, \quad D_s = \begin{pmatrix} 0.8 & 1 & 0 \\ 1 & 0.8 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

$$Q_r = \begin{pmatrix} 10 & 0 & 8 \\ 0 & 10 & 8 \\ 0 & 10 & 0 \end{pmatrix}, \quad q_{\max} = 10, \quad q_{\min} = 0, \quad D_r = \begin{pmatrix} 0 & 1 & 0.2 \\ 1 & 0 & 0.2 \\ 1 & 0 & 1 \end{pmatrix}$$

Appendix C

Statistical content and functional content

To give a semantics is to associate a certain kind of formal object with each element in a system of signs. Intuitively, this formal object captures the content of a sign, what it 'says about the world'.

At least two semantics are relevant for signals and cues in biological systems. They attribute what we shall call **statistical content** and **functional content**. The first is due to Skyrms (2010, §3) and Isaac (2018), while the second is due to Shea et al. (2017).

Statistical content

Consider a probability distribution P over world states W. The statistical content of a sign s, with respect to W, is:

$$\underbrace{S(s)}_{\substack{\text{statistical}\\ \text{content of } s}} \equiv \; < \log \frac{p(w_1|s)}{p(w_1)}, \; ..., \; \underbrace{\log \frac{p(w_i|s)}{p(w_i)}}_{\substack{\text{pointwise mutual}\\ \text{information of}\\ w_i \text{ and } s}}, \; ... \; > \tag{C.1}$$

The statistical content of a sign s is a vector, each of whose components i is the pointwise mutual information between s and w_i. (Isaac (2018) calls such vectors *s-vectors*.) They capture probabilistic relationships between signs and world affairs. A typical statistical content vector looks like this:

$$< 0.5, -4, 3, -\infty >$$

The sign to which this content belongs is statistically associated with four world states, w_1 through w_4. It may or may not be causally associated with them. Each component of the vector measures the change in log probability of the corresponding world state. States 1 and 3 become more likely when the sign occurs. State 2 becomes less likely, and state 4 becomes impossible.

Why did Skyrms choose $\log \frac{p(w_i|s)}{p(w_i)}$ as his vector entries? Because they are the components of **relative entropy**, a useful quantity in information-theoretic applications such as Bayesian epistemology machine learning. Relative entropy measures information gained about hypotheses when evidence is observed. Its general form is:

$$D(P(X)||Q(X)) = \Sigma_x p(x) \log \frac{p(x)}{q(x)}$$

Interpreting Q as the prior probability distribution over hypotheses, $p(H)$, and P the posterior, $p(H|e)$, relative entropy measures the difference between prior and posterior beliefs about hypotheses H in light of evidence e:

$$\underbrace{D(p(H|e)||p(H))}_{\substack{\text{information gained} \\ \text{about } H \text{ from } e}} = \underbrace{\Sigma_h}_{\substack{\text{sum over} \\ \text{hypotheses}}} \underbrace{p(h|e)}_{\substack{\text{posterior} \\ \text{probability} \\ \text{of } h \text{ given } e}} \cdot \underbrace{\log \frac{p(h|e)}{p(h)}}_{\substack{\text{information gained} \\ \text{about } h \text{ from } e}}$$

The idea is that observing evidence enables a more accurate picture of the probabilities of different hypotheses. There are of course many different ways of measuring the difference between two probability distributions $p(H)$ and $p(H|e)$. Relative entropy is special in part because it measures the number of extra bits – binary symbols – one would need to represent the outcomes of events drawn from $p(H|e)$ on the inaccurate assumption that they were

drawn from $p(H)$. In short, it offers a concrete and useful measure of the cost of inaccuracy. Consequently, it can be interpreted as what is *gained* when one learns $p(H|e)$ from a starting position of $p(H)$.

Bringing this back to statistical content, Skyrms interprets H as the hypothesis about which world state obtains, and e as the sign to which the vector belongs. Components of the vector, weighted by posterior probabilities $p(H|e)$, sum to the relative entropy between prior and posterior probabilities over the world states in question. And this list of relative entropies, weighted by the probability of receiving the corresponding signal, sums to the mutual information between signals and states: $I(W;S)$. This places an upper bound on the transmission rate of the signalling system, defined in appendix B.

Functional content

Functional content is also a vector, but its definition is a little more involved. The idea is to capture relationships with world states that contribute to the use of that signal in the same situation. Components of the vector for signal s denote how much greater payoff is obtained with a signal than without. We first need to know the payoffs that would occur in each state if no signalling took place. Let $*$ denote the best strategy in the absence of signalling, so q^r_{i*} denotes the receiver's expected payoff in state i in the absence of signalling.[1] Then we can define each agent's expected payoff in the absence of signalling as follows:

[1] Shea et al. (2017, p. 23) claim only pure strategies need be considered because "in the absence of signals, the receiver can never do better by mixing than by pursuing some pure strategy." This is true for standard implementations of signalling models: a payoff of zero is neutral, rather than a penalty, and lineages cannot go extinct. There are other ways of modelling evolutionary processes, however, in which payoffs denote the number of offspring an agent will have (Donaldson-Matasci et al., 2010). In these models, agents might do better with mixed rather than pure strategies (Kelly, 1956).

$$\overline{q^s} \equiv \Sigma_i p(w_i) q^s_{i*}$$

$$\overline{q^r} \equiv \Sigma_i p(w_i) q^r_{i*}$$

Vector entries are determined by the extent to which payoffs given signals exceed these baseline payoffs. For example, $q^r_{ik} - \overline{q^r}$ is the extent to which doing a_k in state w_i pays the receiver more than doing the act determined by the default strategy.

For the receiver, the expected extra payoff from heeding signals about state w_i is:

$$\underbrace{e^r_i}_{\substack{\text{expected extra benefit}\\ \text{of heeding signals}\\ \text{about state } w_i}} \equiv \underbrace{\Sigma_k}_{\substack{\text{sum over}\\ \text{all acts } a_k}} \underbrace{p(a_k|w_i)}_{\substack{\text{probability of}\\ \text{performing } a_k\\ \text{given } w_i}} \cdot \underbrace{(q^r_{ik} - \overline{q^r})}_{\substack{\text{extra benefit of } a_j\\ \text{over default act}\\ \text{in state } w_i}}$$

The vector component denoting state w_i for signal s is the expected extra benefit of heeding signals about state w_i multiplied by the probability of state w_i given that signal s occurred:

$$\underbrace{x^r_i}_{\substack{\text{vector component}\\ \text{denoting state } w_i\\ \text{for signal } s}} \equiv \begin{cases} \underbrace{p(w_i|s)}_{\substack{\text{probability of state } w_i\\ \text{given signal } s \text{ occurred}}} \cdot \underbrace{e^r_i}_{\substack{\text{expected extra benefit}\\ \text{of heeding signals}\\ \text{about state } w_i}} & \textbf{if } \underbrace{\Sigma_k p(a_k|s) q^r_{ik} > \overline{q^r}}_{\substack{\text{heeding this signal}\\ \text{in this state is}\\ \text{beneficial on average}}} \\ 0 & \textbf{otherwise} \end{cases}$$

Finally, Shea et al. (2017) normalise the vector so that it sums to 1:

$$\sigma = \Sigma_i x^r_i$$

$$F_r(s) \equiv \underbrace{< \frac{x^r_1}{\sigma}, \frac{x^r_2}{\sigma}, \ldots \frac{x^r_i}{\sigma} \ldots >}_{\substack{\text{functional content}\\ \text{of signal } s\\ \text{with respect to receiver}}}$$

This procedure gives different functional contents for sender and receiver. In fact, Shea et al.

(2017, Appendix) define functional content for both agents at once. This requires defining d_{ik}, the minimum increase over baseline payoffs across both agents. Subsequent terms are defined analogously:

$$d_{ik} \equiv \min\left(q^r_{ik} - \overline{q^r},\ q^s_{ik} - \overline{q^s}\right)$$

$$e_i \equiv \Sigma_k p(a_k|w_i).d_{ik}$$

The vector component x_i requires both conditions to be met:

$$x_i \equiv \begin{cases} p(w_i|s).e_i & \textbf{if } \underbrace{\Sigma_k p(a_k|s)q^r_{ik} > \overline{q^r}}_{\substack{\text{heeding this signal}\\ \text{in this state is}\\ \text{beneficial for receiver}}} \textbf{ and } \underbrace{\Sigma_k p(a_k|s)q^s_{ik} > \overline{q^s}}_{\substack{\text{heeding this signal}\\ \text{in this state is}\\ \text{beneficial for sender}}} \\ 0 & \textbf{otherwise} \end{cases}$$

Finally, the vector is normalised and functional content is defined as usual:

$$\sigma \equiv \Sigma_i x_i$$

$$F(s) \equiv < \frac{x_1}{\sigma},\ \frac{x_2}{\sigma},\ \ldots \frac{x_i}{\sigma} \ldots >$$

Functional content captures world states that must obtain in order for receiver behaviour to be successful in responding to a sign. A typical content vector looks like this:

$$< 0, 0, 1, 0 >$$

The signal to which this content belongs is functionally associated with four world states. The third state w_3 is one for which, upon receipt of the signal, acting differently from the default strategy brings greater-than-baseline payoffs.

Appendix D

Grafen's costly signalling model

NB: The variables in this appendix are taken from Grafen (1990), and therefore do not correspond to variables in the rest of the thesis.

As in Grafen's exposition, we will describe the senders as males and receivers as females. The goal of each male is to improve its chances of mating. The goal of each female is to distinguish high-quality from low-quality males – in other words, to accurately assess each male's quality.

The following variables are used (see also Grafen (1990, §2)):

- q: true quality of male
- $a = A(q)$: level of advertising of male; a function of his true quality
- $p = P(a) = P(A(q))$: perceived quality of male; a function of his level of advertising

Given these definitions, we can say that a female perceives 'correctly' when $p = q$. We now write two expressions to capture the fitness of males and females with respect to the interaction mediated by a signal:

- Male fitness: $w = w(a, p, q)$: a function of his level of advertising, actual quality and perceived quality
- Female fitness: $D(q, p)$: a function of the *loss* in fitness suffered by a female who perceives a male of quality q as being of quality p. When $p = q$, the female has perceived correctly, and $D(q, p) = 0$ so there is no fitness penalty.

We need a few more definitions in hand before we can approach the proof:

- $G(q)$: the cumulative frequency distribution of q among males
- $\overline{D}$: Average fitness penalty to females

(Grafen doesn't use the term $\overline{D}$; I have introduced it for convenience.) We can calculate the average fitness penalty $\overline{D}$ in terms of the frequency distribution of males $G(q)$:

$$\overline{D} = \int D[q, P(a)] dG(q)$$

Here $D[q, P(a)]$ is the fitness penalty for a female meeting a male of quality q, where we assume his level of advertising a depends on his quality q. We sum those penalties over all the males in the population, as denoted by the integral sign $\int$ and the term $dG(q)$ (the d tells us to sum over all the members of $G(q)$ i.e. all the males).

In terms of evolutionary games, males play strategy a (advertising at a certain level, given their quality) and females play strategy p (inferring the quality of a male, given their level of advertising). For there to be an evolutionarily stable point in this game, there must be a pair of strategies such that each is optimal given the other. Neither males nor females can unilaterally diverge without incurring a fitness penalty. This situation is denoted by two mathematical expressions. The first states that males cannot improve, the second that females cannot improve:

$$\underbrace{w(A^*(q), P^*(A^*(q)), q)}_{\substack{\text{Male fitness} \\ \text{of playing } A^*}} \quad \underbrace{\geq}_{\substack{\text{is greater than} \\ \text{or equal to}}} \quad \underbrace{w(a, P^*(a), q) \text{ for all } a, q.}_{\substack{\text{male fitness of} \\ \text{any other strategy}}} \tag{D.1}$$

$$\underbrace{\int D[q, P^*(A^*(q))] dG(q)}_{\substack{\text{Female fitness } \textit{penalty} \\ \text{of playing } P^*}} \quad \underbrace{\leq}_{\substack{\text{is less than} \\ \text{or equal to}}} \quad \underbrace{\int D[q, P(A^*(q))] dG(q) \text{ for all } P(a).}_{\substack{\text{female fitness } \textit{penalty} \\ \text{of any other strategy}}} \tag{D.2}$$

(These two inequalities are Grafen's equation (1), page 519.) When these inequalities hold, the pair of strategies A^* and P^* are evolutionarily stable.

Another bit of terminology is required:

- w_1, w_2, w_3: partial derivatives of $w(a,p,q)$ with respect to a, p and q respectively.
- w_{11}, w_{12}, w_{21} etc: higher order derivatives.

The statements to be proved

Grafen proves two statements, which we can think of as minor and major handicap results. The first makes use of the notion of an evolutionarily stable strategy (ESS) (Maynard Smith and Price, 1973).

> *Minor handicap result*: "under weak conditions on the function $w(a,p,q)$, an ESS exists which exhibits the features Zahavi associated with the handicap principle" (Grafen, 1990, p. 519)
> *Major handicap result*: "one can conclude from the evolutionary stability of signals that they are honest, costly and costly in a way that relates to the true quality revealed" (Grafen, 1990, p. 521).

Minor result: proof

(See also Grafen (1990, §2.1).)

A few assumptions:

1. The set of points of increase of G is an interval (i.e. there are no gaps in the distribution of quality).
2. w is continuous.
3. w_1, w_2, w_3 exist.

4. w_1 is negative. This means that w goes down as a goes up, which embodies the assumption that advertising is costly.
5. w_2 is positive. This means that w goes up as p goes up, which embodies the assumption that male fitness improves as female perception of their quality increases.
6. $\frac{w_1(a,p,q)}{w_2(a,p,q)}$ is strictly increasing in q. This means that better males do better by advertising more. The condition is satisfied when $w_{23} \geq 0$ and $w_{13} > 0$.
7. The ratio $\frac{w_1}{w_2}$ is defined. This is required for technical reasons.
8. The set of values of q for which $G(q)$ is increasing is assumed to be an interval on the real line, $[q_{\min}, q_{\max}]$ where $q_{\min}$ is finite.
9. There is a finite minimum level of advertising, $a_{\min}$.

Now, functions P^* and A^* can be defined as follows:

$$\begin{aligned} P^*(a_{\min}) &= q_{\min} \\ P^{*\prime}(a) &= -\frac{w_1[a, P^*(a), P^*(a)]}{w_2[a, P^*(a), P^*(a)]} \\ P^*[A^*(q)] &= q \end{aligned} \tag{D.3}$$

P^* is monotone increasing because w_1 and w_2 are of opposite sign (by assumptions 4 and 5).

The marginal value of advertising, given females are adopting strategy P^*, is:

$$\underbrace{\frac{\delta}{\delta a} w[a, P^*(a), q]}_{\text{Marginal value of advertising}} = \underbrace{w_1[a, P^*(a), q]}_{\text{Advertising expenditure}} + \underbrace{P^{*\prime}(a)}_{\substack{\text{female perception} \\ \text{given advertising level}}} \underbrace{w_2[a, P^*(a), q]}_{\substack{\text{fitness benefit} \\ \text{given female perception}}}$$

When the marginal value of advertising is positive, males are incentivised to advertise more. When it is negative, they are incentivised to advertise less. So a stable point will occur

when the marginal value of advertising is zero. We can find the *sign* of the marginal value of advertising – whether it is positive, negative or zero – by doing the following. We substitute in the definition of $P^{*'}$ from equation D.3, and divide by $w_2[a, P^*(a), q]$, so the right hand side becomes:

$$\underbrace{\frac{w_1[a, P^*(a), q]}{w_2[a, P^*(a), q]}}_{\textit{Left hand quotient}} - \frac{w_1[a, P^*(a), P^*(a)]}{w_2[a, P^*(a), P^*(a)]}$$

The marginal value of advertising has the same sign as this term, because all we did was divide by $w_2[a, P^*(a), q]$ which is positive (by assumption 5).

By assumption 6 (better males do better by advertising more), the left hand quotient is increasing in q. Therefore, the marginal value of advertising is positive when $q > P^*(a)$, zero when $q = P^*(a)$ and negative when $q < P^*(a)$. In other words, males are incentivised to advertise more when females perceive them as lower quality than they really are. Inversely, they are incentivised to advertise *less* when females perceive them as *higher* quality than they really are – because in order to achieve it they must expend more resources than a favourable assessment can recoup.

Now, consider the relationship between P^* and A^* described in equation D.3. $P^*(a)$ is an increasing function with inverse A^*, since $P^*[A^*(q)] = q$). Therefore, the marginal value of advertising is positive when $a < A^*(q)$, zero when $a = A^*(q)$ and negative when $a > A^*(q)$. In other words, males are incentivised to advertise more when they are not yet signalling at the 'standard level' A^*. Inversely, they are incentivised to advertise *less* when they are signalling *above* the level of A^*. As a result, A^* is an ESS when all females employ the strategy P^*.

It remains only to show that P^* is an ESS when all males employ the strategy A^*. Recall from equation D.2 that P^* has to satisfy an inequality to be an ESS. Both sides of the inequality can be no lower than zero, so the female strategy will satisfy the condition when the left hand side of the inequality is zero. But we defined the female strategy such that $P^*(A^*(q)) = q$, which entails that the left hand side of the inequality is always zero. Therefore it achieves the minimum possible fitness penalty, and is an ESS.

Therefore, P^* and A^* are an evolutionarily stable pair of strategies.

Major result: proof

(See also Grafen (1990, §2.2).)

In formal terms, the statement to be proved is the following:

If A^*, P^* is an ESS pair, $w_2 > 0$ and $A^*(q)$ is increasing, then

(a) $P^*[A^*(q)] = q \qquad \forall q$

(b) $w_1 < 0$

(c) $\dfrac{w_1(a,p,q)}{w_2(a,p,q)}$ is strictly increasing in q near the path $[A^*(q), q, q]$

Condition (a) corresponds to honesty of signals, condition (b) embodies the costliness of signals, and condition (c) captures the idea that signals should be costlier for males of lower quality. The proof proceeds in three parts.

(a). Since $A(q)$ is increasing, there is an inverse function A^{-1} mapping each level of advertising into the quality of the male producing it. Female fitness penalty is minimised when $P = A^{-1}$, so females will play this strategy. Therefore, $P(A(q)) = A^{-1}(A(q)) = q$ and (a) holds. In other words, if signals were not reliable, females would not use them. (This

result assumes a free choice of female strategy.)

(b). $A(q)$ is the male strategy, and it is stable when it maximises male fitness $w(a,p,q)$. We saw that the following must be true when w is maximised:

$$w_1 + P'w_2 = 0 \tag{D.4}$$

But we know that P' is positive, because A' is. w_2 is positive by assumption, therefore w_1 must be negative and (b) holds. In other words, if signalling is stable and honest, there must be a reason why worse males do not signal at a higher level. This reason can only be cost. (This result assumes a free choice of male strategy.)

(c). Equation D.4 entails that male fitness is at either a maximum or a minimum. For it to be a maximum, as per the assumption that it is stable, (c) must hold. In other words, each male's level of advertising is stable (i.e. provides maximum fitness), yet better males advertise more. Therefore the marginal cost of advertising must be lower for better males.

Appendix E

de Polavieja's costly signalling model

The following model is adapted from de Polavieja (2002). Some variable names have been changed for consistency.

Basics. Let $W = (w_1, w_2, ...w_N)$ be states of the world, $S = (s_1, s_2, ...s_N)$ signals. For simplicity, assume there are the same number of signals as states, and the receiver must infer the state from the signal.[1]

Error. Let E be a matrix such that $e_{ij} = p(w_i|s_j)$, the conditional probability that w_i is the state when signal s_j was sent. Call this the **error matrix**. When there are no errors, every entry except the diagonal is zero (it is a *diagonal matrix*). When there are errors, the matrix is not diagonal: some states prompt signals that do not correspond to them. Conversely, let F be a matrix such that $f_{ji} = p(s_j|w_i)$, the conditional probability that s_j is emitted when w_i is the state.

Cost. Let c_{ij} be the cost of producing signal s_j in state w_i. The average cost of signal s_j is $c_j = \Sigma_i e_{ij} c_{ij}$. The average cost across all signals is $\bar{c} = \Sigma_j p(s_j) c_j$. Index the signals in order of increasing cost, so $c_1 \leq c_2 \leq ... \leq c_N$.

Information. Mutual information between states W and signals S is:

$$I(W;S) = \Sigma_{i,j} p(w_i, s_j) \log \frac{p(w_i, s_j)}{p(w_i)p(s_j)}$$

[1] Two things I haven't been able to confirm: whether states are equiprobable, and whether there could be a different number of signals than states. I believe the answer to be yes to both.

Now, the entropy of signals $H(S)$ is defined as

$$H(S) = \Sigma_j p(s_j) \log \frac{1}{p(s_j)}$$

And if we define **signal uncertainty** u_j for signal s_j as

$$u_j = \Sigma_i e_{ij} \log \frac{1}{f_{ji}} \tag{E.1}$$

then the **noise entropy** (conditional entropy of signals given states) $H(S|W)$ is

$$H(S|W) = \Sigma_j p(s_j) u_j$$

Putting together these equations, we can express mutual information between signals and states in terms of the signal variability and error constraint:

$$\begin{aligned} I(W;S) &= H(S) - H(S|W) \\ &= \underbrace{\Sigma_j p(s_j) \log \frac{1}{p(s_j)}}_{\text{signal variability}} - \underbrace{\Sigma_j p(s_j) u_j}_{\text{error constraint}} \end{aligned}$$

What is going on here? Usually when considering the mutual information between two things when one is downstream of the other, we take the *upstream* one as the main variable. $H(W) - H(W|S)$ means something like 'what it's possible to learn about w, minus what you still don't know after observing s'. Unusually, here the equation is the other way round. It still equals the same value, but the natural interpretation is slightly different. $H(S) - H(S|W)$ means something like 'what it's possible to learn from s, minus what it refuses to tell you'. In the present model, we are interpreting 'refuses to tell you' as 'can*not* tell you, due to

inherent error'. As such, we defined $H(S|W)$ solely in terms of the error matrix E. There was an f_{ji} term in equation E.1, but that can be rewritten in terms of e_{ij} by Bayes' theorem:

$$f_{ji} = \frac{p(s_j)e_{ij}}{\Sigma_k p(s_k)e_{ik}}$$

The problem to be solved. We have a definition of transmission rate and a term that represents the average cost of a signal. We want to maximise transmission rate $I(W;S)$ while not exceeding a certain average cost $c_{\max}$. This can be done by altering the frequency with which signals are used – that is, altering the probability distribution over signals, $P(S)$. Notice an important difference with the way signalling games are traditionally represented: usually the sender strategy matrix $P(S|W)$ determines the conditional probabilities of signals given states and is interpreted as the sender's choice; here the converse error matrix F determines the conditional probabilities of signals given states and is interpreted as a consequence of noise. The sender's assignment of meanings to signals is determined by the choice of signal frequencies $P(S)$.

This seems like a problem that evolution by natural selection might plausibly be interpreted as solving. de Polavieja's example is neural signalling. Neurons need to communicate, but it costs them energy to do so. Selection produces signalling strategies that must trade off metabolic cost for accuracy. What would the best strategy be?

The solution. Here's where things get a bit involved. When maximising a function subject to constraints, we use Lagrange multipliers (Jensen, 2006). A variation on the Blahut-Arimoto algorithm (Arimoto, 1972; Blahut, 1972) allows us to calculate the maximum value of $I(W;S)$ such that $\bar{c} \leq c_{\max}$. The algorithm runs as follows.

Step 1. Pick a random starting vector $\hat{p}(S)$.

Step 2. Using $\hat{p}(S)$, calculate the conditional matrix f_{ji}:

$$f_{ji} = \frac{p(s_j)e_{ij}}{\Sigma_k p(s_k)e_{ik}} \tag{E.2}$$

Step 3. Use the conditional matrix to calculate the cost constraint β from the equation:[2]

$$c_{\max} = \frac{\Sigma_j c_j \exp -\left(\beta c_j - \Sigma_k e_{jk} \log f_{jk}\right)}{\Sigma_j \exp -\left(\beta c_j - \Sigma_k e_{jk} \log f_{jk}\right)} \tag{E.3}$$

Step 4. Use β to calculate the new vector $\hat{p}(S)$:

$$\hat{p}(S) = \frac{\exp -\left(\beta c_j - \Sigma_k e_{kj} \log f_{jk}\right)}{\Sigma_i \exp -\left(\beta c_i - \Sigma_k e_{ki} \log f_{ik}\right)} \tag{E.4}$$

Step 5. Repeat steps 2-4 (equations E.2, E.3 and E.4, in order) until $\hat{p}(S)$ no longer changes. The value to which $\hat{p}(S)$ converges is the optimal usage of signals.

Example. Consider a system with three signals such that the first two cost the same, $c_1 = c_2$, but the third costs more, $c_3 > c_1$. Suppose the error matrix is:

$$E = p(w_i|s_j) = \begin{pmatrix} 1-\rho & \rho & 0 \\ \rho & 1-\rho & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The noise entropy is:

$$H(S|W) = (p(s_1) + p(s_2))H_2(\rho) \tag{E.5}$$

Where $H_2(\rho)$ is the binary entropy function $\rho \log \frac{1}{\rho} + (1-\rho) \log \frac{1}{1-\rho}$.

[2] It is not clear to me how to calculate β, because it is not easy to rearrange this equation in order to isolate β on one side. But as it is a necessary part of the algorithm, I include it here for completeness.

The algorithm delivers the following probabilities for signals (de Polavieja, 2002, p. 663):

$$p(s_1) = p(s_2) = \frac{\exp(-\beta c_1 - H_2(\rho))}{Z} \tag{E.6}$$

$$p(s_3) = \frac{\exp(-\beta c_3)}{Z} \tag{E.7}$$

Where Z is a normalisation constant required to ensure $p(s_1) + p(s_2) + p(s_3) = 1$:

$$Z = 2\exp(-\beta c_1 - H_2(\rho)) + \exp(-\beta c_3)$$

And β is determined by taking the average energy to be the allowed maximum, in other words substituting equations E.6 and E.7 into equation E.8 and solving for β:

$$c_{\max} = 2p(s_1)c_1 + p(s_3)c_3 \tag{E.8}$$

The equations are tricky, but inspecting E.6 and E.7 reveals an intuitive result. Signals s_1 and s_2 are at risk of being conflated, and so their use is penalised by $H_2(\rho)$: the frequency with which they ought to be used is inversely proportional to the extent they are liable to be mistaken for each other. Signal s_3 cannot be confused with any other, so the only thing that penalises its use is the cost c_3. Despite the lower cost $c_1 = c_2$ of the cheaper signals, the need to reliably distinguish states might lead to s_3 being used more; it depends on the relative sizes of βc_3 and $\beta c_1 + H_2(\rho)$.

Bibliography

Alexander, Richard D. and Gerald Borgia (1978). "Group Selection, Altruism, and the Levels of Organization of Life". *Annual Review of Ecology and Systematics* 9.1, pp. 449–474. Link.

Arimoto, Suguru (1972). "An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels". *IEEE Transactions on Information Theory* 18.1, pp. 14–20. Link.

Artiga, Marc (2014a). "Signaling without Cooperation". *Biology & Philosophy* 29.3, pp. 357–378. Link.

— (2014b). "Teleosemantics, Infotel-Semantics and Circularity". *International Journal of Philosophical Studies* 22.4, pp. 583–603. Link.

— (2016a). "Liberal Representationalism: A Deflationist Defense". *Dialectica* 70.3, pp. 407–430. Link.

— (2016b). "Teleosemantic Modeling of Cognitive Representations". *Biology & Philosophy* 31.4, pp. 483–505. Link.

— (2020). "Signals Are Minimal Causes". *Synthese*. Link.

Attneave, Fred (1954). "Some Informational Aspects of Visual Perception". *Psychological Review* 61.3, pp. 183–193. Link.

Bar-Hillel, Yehoshua and Rudolf Carnap (1953). "Semantic Information". *The British Journal for the Philosophy of Science* 4.14, pp. 147–157. Link.

Beekman, Madeleine et al. (2008). "Dance Precision of Apis Florea—Clues to the Evolution of the Honeybee Dance Language?" *Behavioral Ecology and Sociobiology* 62.8, pp. 1259–1265. Link.

Beekman, Madeleine et al. (2015). "Honeybee Linguistics—a Comparative Analysis of the Waggle Dance among Species of Apis". *Frontiers in Ecology and Evolution* 3, p. 11. Link.

Bentz, Christian (2018). *Adaptive Languages: An Information-Theoretic Account of Linguistic Diversity*. Walter de Gruyter GmbH.

Berger, T., Zhen Zhang, and H. Viswanathan (1996). "The CEO Problem [Multiterminal Source Coding]". *IEEE Transactions on Information Theory* 42.3, pp. 887–902. Link.

Bergstrom, Carl T. and Michael Lachmann (1998). "Signaling among Relatives. III. Talk Is Cheap". *Proceedings of the National Academy of Sciences* 95.9, pp. 5100–5105. Link.

— (2003). "The Red King Effect: When the Slowest Runner Wins the Coevolutionary Race". *Proceedings of the National Academy of Sciences* 100.2, pp. 593–598. Link.

— (2004). "Shannon Information and Biological Fitness". *Information Theory Workshop, 2004*. IEEE, pp. 50–54. Link.

Bergstrom, Carl T. and Martin Rosvall (2011). "The Transmission Sense of Information". *Biology & Philosophy* 26.2, pp. 159–176. Link.

Biesmeijer, Jacobus C. and Thomas D. Seeley (2005). "The Use of Waggle Dance Information by Honey Bees throughout Their Foraging Careers". *Behavioral Ecology and Sociobiology* 59.1, pp. 133–142. Link.

Birch, Jonathan (2014a). "Hamilton's Rule and Its Discontents". *The British Journal for the Philosophy of Science* 65.2, pp. 381–411. Link.

Birch, Jonathan (2014b). “Propositional Content in Signalling Systems”. *Philosophical Studies* 171.3, pp. 493–512. Link.

— (2017). *The Philosophy of Social Evolution*. Oxford University Press.

Blahut, R. (1972). “Computation of Channel Capacity and Rate-Distortion Functions”. *IEEE Transactions on Information Theory* 18.4, pp. 460–473. Link.

Bourke, Andrew F. G. (2014). “Hamilton’s Rule and the Causes of Social Evolution”. *Phil. Trans. R. Soc. B* 369.1642, p. 10. Link.

Bruce, John B. et al. (2017). “Cheating and Resistance to Cheating in Natural Populations of the Bacterium Pseudomonas Fluorescens”. *Evolution* 71.10, pp. 2484–2495. Link.

Bruner, Justin P., Carl Brusse, and David Kalkman (2017). “Cost, Expenditure and Vulnerability”. *Biology & Philosophy* 32.3, pp. 357–375. Link.

Brusse, Carl and Justin Bruner (2017). “Responsiveness and Robustness in the David Lewis Signaling Game”. *Philosophy of Science* 84.5, pp. 1068–1079. Link.

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

Cabirol, Amélie et al. (2018). “Relationship between Brain Plasticity, Learning and Foraging Performance in Honey Bees”. *PLOS ONE* 13.4, e0196749. Link.

Cao, Rosa (2012). “A Teleosemantic Approach to Information in the Brain”. *Biology & Philosophy* 27.1, pp. 49–71. Link.

— (2014). “Signaling in the Brain: In Search of Functional Units”. *Philosophy of Science* 81.5, pp. 891–901. Link.

Chemero, Anthony (2009). *Radical Embodied Cognitive Science*. MIT Press.

Chittka, Lars (2004). “Dances as Windows into Insect Perception”. *PLOS Biology* 2.7, pp. 898–900. Link.

Churchland, Paul M. (1996). *The Engine of Reason, the Seat of the Soul: A Philosophical Journey Into the Brain*. MIT Press.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. Second. Hoboken, New Jersey: John Wiley & Sons.

Craver, Carl F. and William Bechtel (2007). "Top-down Causation Without Top-down Causes". *Biology & Philosophy* 22.4, pp. 547–563. Link.

Craver, Carl F. and David M. Kaplan (2018). "Are More Details Better? On the Norms of Completeness for Mechanistic Explanations". *The British Journal for the Philosophy of Science*. Link.

Cummings, Molly E. (2007). "Sensory Trade-Offs Predict Signal Divergence in Surfperch". *Evolution* 61.3, pp. 530–545. Link.

Cummins, Robert (1996). *Representations, Targets, and Attitudes*. MIT Press.

Dall, Sasha R. X. et al. (2005). "Defining the Concept of Public Information". *Science* 308.5720, pp. 353–356. Link.

Dawkins, Marian Stamp (1993). "Are There General Principles of Signal Design?" *Philosophical Transactions: Biological Sciences* 340.1292, pp. 251–255. Link.

Dawkins, Richard and John R. Krebs (1978). "Animal Signals: Information or Manipulation?" *Behavioural Ecology: An Evolutionary Approach*. Ed. by J. R. Krebs and N. B. Davies. First, pp. 282–309.

de Polavieja, Gonzalo G. (2002). "Errors Drive the Evolution of Biological Signalling to Costly Codes". *Journal of Theoretical Biology* 214.4, pp. 657–664. Link.

Dennett, Daniel C. (1983). "Intentional Systems in Cognitive Ethology: The "Panglossian Paradigm" Defended". *Behavioral and Brain Sciences* 6.3, pp. 343–355. Link.

Dennett, Daniel C. (1987). *The Intentional Stance*. MIT Press.

— (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Penguin UK.

Diggle, Stephen P. et al. (2007a). "Cooperation and Conflict in Quorum-Sensing Bacterial Populations". *Nature* 450.7168, pp. 411–414. Link.

Diggle, Stephen P. et al. (2007b). "Evolutionary Theory of Bacterial Quorum Sensing: When Is a Signal Not a Signal?" *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362.1483, pp. 1241–1249. Link.

Donaldson-Matasci, Matina C., Carl T. Bergstrom, and Michael Lachmann (2010). "The Fitness Value of Information". *Oikos* 119.2, pp. 219–230. Link.

Dretske, Fred (1981). *Knowledge and the Flow of Information*. Cambridge, Mass.: MIT Press.

— (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, Mass.: MIT Press.

Dronamraju, Krishna R. (2017). *Popularizing Science: The Life and Work of JBS Haldane*. New York: Oxford University Press.

Dumont, Elizabeth R. et al. (2014). "Selection for Mechanical Advantage Underlies Multiple Cranial Optima in New World Leaf-Nosed Bats". *Evolution* 68.5, pp. 1436–1449. Link.

Dyson, Freeman (2011). "How We Know". Link.

El Gamal, Abbas and Young-Han Kim (2011). *Network Information Theory*. Cambridge University Press. Link.

Endler, John A. (1993). "Some General Comments on the Evolution and Design of Animal Communication Systems". *Philosophical Transactions: Biological Sciences* 340.1292, pp. 215–225. Link.

Fodor, Jerry A. (1990). *A Theory of Content and Other Essays*. Cambridge, Mass.: MIT Press.

Forterre, Yoël et al. (2005). "How the Venus Flytrap Snaps". *Nature* 433.7024, p. 421. Link.

Frank, Steven A. (1998). *Foundations of Social Evolution*. Princeton University Press.

Fraser, Ben (2012). "Costly Signalling Theories: Beyond the Handicap Principle". *Biology & Philosophy* 27.2, pp. 263–278. Link.

Fresco, Nir, Simona Ginsburg, and Eva Jablonka (2018). "Functional Information: A Graded Taxonomy of Difference Makers". *Review of Philosophy and Psychology*, pp. 1–21. Link.

Fuchs, Gilad et al. (2014). "4sUDRB-Seq: Measuring Genomewide Transcriptional Elongation Rates and Initiation Frequencies within Cells". *Genome Biology* 15.5, R69. Link.

Gallistel, Charles R. (2020). "Where Meanings Arise and How: Building on Shannon's Foundations". *Mind & Language* 35.3, pp. 390–401. Link.

Gärdenfors, Peter (2004). "Conceptual Spaces as a Framework for Knowledge Representation". *Mind and Matter* 2.2, pp. 9–27. Link.

— (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Gelder, Tim van (1998). "The Dynamical Hypothesis in Cognitive Science". *Behavioral and Brain Sciences* 21.5, pp. 615–628. Link.

Gibson, Thomas C. and Donald M. Waller (2009). "Evolving Darwin's 'Most Wonderful' Plant: Ecological Steps to a Snap-Trap". *New Phytologist* 183.3, pp. 575–587. Link.

Gładziejewski, Paweł and Marcin Miłkowski (2017). "Structural Representations: Causally Relevant and Different from Detectors". *Biology & Philosophy* 32.3, pp. 337–355. Link.

Godfrey-Smith, Peter (1996). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Godfrey-Smith, Peter (2009). *Darwinian Populations and Natural Selection*. Oxford: OUP.

— (2011). "Signals: Evolution, Learning, and Information, by Brian Skyrms." *Mind* 120.480, pp. 1288–1297. Link.

Godfrey-Smith, Peter and Manolo Martínez (2013). "Communication and Common Interest". *PLOS Computational Biology* 9.11, e1003282. Link.

Godfrey-Smith, Peter and Kim Sterelny (2016). "Biological Information". *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2016. Link.

Gorrell, Jamieson C. et al. (2010). "Adopting Kin Enhances Inclusive Fitness in Asocial Red Squirrels". *Nature Communications* 1, p. 22. Link.

Gould, J. L. (1975). "Honey Bee Recruitment: The Dance-Language Controversy". *Science* 189.4204, pp. 685–693. Link.

Gould, S. J. and R. C. Lewontin (1979). "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme". *Proceedings of the Royal Society of London. Series B. Biological Sciences* 205.1161, pp. 581–598. Link.

Grafen, Alan (1990). "Biological Signals as Handicaps". *Journal of Theoretical Biology* 144.4, pp. 517–546. Link.

Grice, H. P. (1957). "Meaning". *The Philosophical Review* 66.3, pp. 377–388. Link.

Griffiths, Paul E. (2018). *What Distinctions Can Natural Selection Draw?* University of Wollongong. Link.

Griffiths, Paul E. et al. (2015). "Measuring Causal Specificity". *Philosophy of science* 82.4, pp. 529–555. Link.

Grüter, Christoph, M. Sol Balbuena, and Walter M. Farina (2008). "Informational Conflicts Created by the Waggle Dance". *Proceedings of the Royal Society of London B: Biological Sciences* 275.1640, pp. 1321–1327. Link.

Grüter, Christoph and Francis L. W. Ratnieks (2011). "Honeybee Foragers Increase the Use of Waggle Dance Information When Private Information Becomes Unrewarding". *Animal Behaviour* 81.5, pp. 949–954. Link.

Guilford, Tim and Marian Stamp Dawkins (1991). "Receiver Psychology and the Evolution of Animal Signals". *Animal Behaviour* 42.1, pp. 1–14. Link.

Gupta, Shreejata and Anindya Sinha (2019). "Gestural Communication of Wild Bonnet Macaques in the Bandipur National Park, Southern India". *bioRxiv*, p. 739441. Link.

Hailman, Jack Parker (2008). *Coding and Redundancy: Man-Made and Animal-Evolved Signals*. Cambridge, Mass.: Harvard University Press.

Haldane, J. B. S. and H. Spurway (1954). "A Statistical Analysis of Communication in "Apis Mellifera" and a Comparison with Communication in Other Animals". *Insectes Sociaux* 1.3, pp. 247–283. Link.

Hamilton, W. D. (1964a). "The Genetical Evolution of Social Behaviour. I". *Journal of Theoretical Biology* 7.1, pp. 1–16. Link.

— (1964b). "The Genetical Evolution of Social Behaviour. II". *Journal of Theoretical Biology* 7.1, pp. 17–52. Link.

Hanlon, Roger T. et al. (2005). "Transient Sexual Mimicry Leads to Fertilization". *Nature* 433.7023, pp. 212–212. Link.

Harms, William F. (2004). "Primitive Content, Translation, and the Emergence of Meaning in Animal Communication". *Evolution of Communication Systems: A Comparative Ap-*

proach. Ed. by D. Kimbrough Oller and Ulrike Griebel. Cambridge, Mass.: MIT Press, pp. 31–48.

Hartley, R. V. L. (1928). "Transmission of Information". *Bell System Technical Journal* 7.3, pp. 535–563. Link.

Haugeland, John (1998). *Having Thought*. Harvard University Press.

Healy, Susan D and Candy Rowe (2007). "A Critique of Comparative Studies of Brain Size". *Proceedings of the Royal Society B: Biological Sciences* 274.1609, pp. 453–464. Link.

Hofstadter, Douglas (1985). *Metamagical Themas: Questing For The Essence Of Mind And Pattern*. Basic Books.

Hofstadter, Douglas R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.

Hughes, David T. and Vanessa Sperandio (2008). "Inter-Kingdom Signalling: Communication between Bacteria and Their Hosts". *Nature Reviews Microbiology* 6.2, pp. 111–120. Link.

Hurd, Peter L. and Magnus Enquist (2005). "A Strategic Taxonomy of Biological Communication". *Animal Behaviour* 70.5, pp. 1155–1170. Link.

Hutchinson, G. Evelyn (1957). "Concluding Remarks". *Cold Spring Harbor Symposia on Quantitative Biology* 22, pp. 415–427. Link.

Hutto, Daniel D. and Erik Myin (2013). *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, Mass.: MIT Press.

I'Anson Price, R. et al. (2019). "Honeybees Forage More Successfully without the "Dance Language" in Challenging Environments". *Science Advances* 5.2, eaat0450. Link.

I'Anson Price, Robbie and Christoph Grüter (2015). "Why, When and Where Did Honey Bee Dance Communication Evolve?" *Frontiers in Ecology and Evolution* 3, p. 125. Link.

Iglesias, P. A. (2016). “The Use of Rate Distortion Theory to Evaluate Biological Signaling Pathways”. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* 2.1, pp. 31–39. Link.

Isaac, Alistair M. C. (2018). “The Semantics Latent in Shannon Information”. *The British Journal for the Philosophy of Science*. Link.

Jacob, François (1973). *The Logic of Life*. Trans. by Betty E. Spillman. Princeton University Press.

Jäger, Gerhard (2018). “Global-Scale Phylogenetic Linguistic Inference from Lexical Resources”. *Scientific Data* 5, p. 180189. Link.

Jeffrey, Richard C. (1990). *The Logic of Decision*. University of Chicago Press.

Jensen, Steuard (2006). *An Introduction to Lagrange Multipliers*. Link.

Jimbo, Masakazu and Kiyonori Kunisawa (1979). “An Iteration Method for Calculating the Relative Capacity”. *Information and Control* 43.2, pp. 216–223. Link.

Kanwal, Jasmeen et al. (2017). “Zipf’s Law of Abbreviation and the Principle of Least Effort: Language Users Optimise a Miniature Lexicon for Efficient Communication”. *Cognition* 165, pp. 45–52. Link.

Karolinska Institutet (1973). *The Nobel Prize in Physiology or Medicine 1973*. Press Release. Link.

Kelly, J. L. (1956). “A New Interpretation of Information Rate”. *Bell System Technical Journal* 35.4, pp. 917–926. Link.

Kolchinsky, Artemy and David H. Wolpert (2018). “Semantic Information, Autonomous Agency and Non-Equilibrium Statistical Physics”. *Interface Focus* 8.6, p. 20180041. Link.

Krebs, John R. and Nicholas B. Davies, eds. (1978). *Behavioural Ecology: An Evolutionary Approach*. First. Oxford: Blackwell Scientific Publications.

— (1984). *Behavioural Ecology: An Evolutionary Approach*. Second. Blackwell.

Krebs, John R. and Richard Dawkins (1984). "Animal Signals: Mind-Reading and Manipulation". *Behavioural Ecology: An Evolutionary Approach*. Ed. by J. R. Krebs and N. B. Davies. Second. Oxford: Blackwell Scientific, pp. 380–402.

Kripke, Saul A. (1982). *Wittgenstein on Rules and Private Language*. Harvard University Press.

Leal, Manuel (1999). "Honest Signalling during Prey–Predator Interactions in the Lizard Anolis Cristatellus". *Animal Behaviour* 58.3, pp. 521–526. Link.

Lean, Oliver M. (2014). "Getting the Most out of Shannon Information". *Biology & Philosophy* 29.3, pp. 395–413. Link.

Lehmann, Kenna D. S. et al. (2014). "From Cues to Signals: Evolution of Interspecific Communication via Aposematism and Mimicry in a Predator-Prey System". *PLOS ONE* 9.3, e91783. Link.

Lehmann, Laurent and François Rousset (2014). "Fitness, Inclusive Fitness, and Optimization". *Biology & Philosophy* 29.2, pp. 181–195. Link.

Levy, Arnon (2011). "Information in Biology: A Fictionalist Account". *Noûs* 45.4, pp. 640–657. Link.

Lewis, David (1969). *Convention: A Philosophical Study*. Oxford: Blackwell.

Lloyd, J. E. (1975). "Aggressive Mimicry in Photuris Fireflies: Signal Repertoires by Femmes Fatales". *Science (New York, N.Y.)* 187.4175, pp. 452–453. Link.

Machamer, Peter, Lindley Darden, and Carl F. Craver (2000). “Thinking about Mechanisms”. *Philosophy of Science* 67.1, pp. 1–25. Link.

MacKay, D. M. and W. S. McCulloch (1952). “The Limiting Information Capacity of a Neuronal Link”. *The Bulletin of Mathematical Biophysics* 14.2, pp. 127–135. Link.

MacKay, David JC (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. Link.

MacKay, Donald MacCrimmon (1969). *Information, Mechanism and Meaning*. Cambridge, Mass.: M.I.T. Press.

Mann, Stephen Francis (2018). “Attribution of Information in Animal Interaction”. *Biological Theory* 13.3, pp. 164–179. Link.

Margalef, Ramón (1957). “Information Theory in Ecology”. *Memorias de La Real Academia de Ciencias y Artes de Barcelona*. Trans. by Wendell Hall. Real Academia de Ciencias y Artes de Barcelona, pp. 373–449.

Marler, P. (1955). “Characteristics of Some Animal Calls”. *Nature* 176.4470, pp. 6–8. Link.

Marsh, Geoff (2015). “Darwin’s Iconic Finches Join Genome Club”. *Nature News* 518.7538, p. 147. Link.

Martínez, Manolo (2015). “Deception in Sender-Receiver Games”. *Erkenntnis* 80.1, pp. 215–227. Link.

— (2019). “Deception as Cooperation”. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, p. 101184. Link.

— (2020). *Information Theory Is a Formal Theory of Representation*. Link.

Martínez, Manolo and Peter Godfrey-Smith (2016). "Common Interest and Signaling Games: A Dynamic Analysis". *Philosophy of Science* 83.3, pp. 371–392. Link.

Maynard Smith, J. and G. R. Price (1973). "The Logic of Animal Conflict". *Nature* 246.5427, pp. 15–18. Link.

Maynard Smith, John (1982). *Evolution and the Theory of Games*. Cambridge University Press.

Maynard Smith, John and David Harper (2003). *Animal Signals*. OUP Oxford.

McWhirter, Gregory (2016). "Behavioural Deception and Formal Models of Communication". *The British Journal for the Philosophy of Science* 67.3, pp. 757–780. Link.

Mehta, Pankaj et al. (2009). "Information Processing and Signal Integration in Bacterial Quorum Sensing". *Molecular Systems Biology* 5, p. 325. Link.

Meyer, Russell (2018). "The Non-Mechanistic Option: Defending Dynamical Explanations". *The British Journal for the Philosophy of Science*, pp. 1–27. Link.

Millikan, Ruth G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press. Link.

— (1989). "Biosemantics". *The Journal of Philosophy* 86.6, pp. 281–297. Link.

— (1993). *White Queen Psychology and Other Essays for Alice*. Bradford Books.

— (2001). "What Has Natural Information to Do with Intentional Representation?" *Royal Institute of Philosophy Supplements* 49, pp. 105–125. Link.

— (2004a). "On Reading Signs: Some Differences between Us and the Others". *Evolution of Communication Systems: A Comparative Approach*. Ed. by D. Kimbrough Oller and Ulrike Griebel, pp. 15–29. Link.

— (2004b). *Varieties of Meaning*. Cambridge, Mass.: MIT Press. Link.

— (2007). “An Input Condition for Teleosemantics? Reply to Shea (and Godfrey-Smith)”. *Philosophy and Phenomenological Research* 75.2, pp. 436–455. Link.

— (2013a). “Natural Information, Intentional Signs and Animal Communication”. *Animal Communication Theory*. Ed. by Ulrich E. Stegmann. New York: Cambridge University Press, pp. 133–146. Link.

— (2013b). “Reply to Rescorla”. *Millikan and Her Critics*. Ed. by Dan Ryder, Justine Kingsbury, and Kenneth Williford. John Wiley & Sons, pp. 103–106.

— (2017). *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford: OUP.

Millikan, Ruth Garrett (2013c). “Reply to Shea”. *Millikan and Her Critics*. Ed. by Dan Ryder, Justine Kingsbury, and Kenneth Williford. John Wiley & Sons, pp. 81–86.

— (2018a). “Representation Made Simple: The Irrelevance of Intensionality”.

— (2018b). *Representations Made Simple*. Link.

Mitchell, Sandra D. (2000). “Dimensions of Scientific Law”. *Philosophy of Science* 67.2, pp. 242–265. Link.

Montealegre-Z, F. (2009). “Scale Effects and Constraints for Sound Production in Katydids (Orthoptera: Tettigoniidae): Correlated Evolution between Morphology and Signal Parameters”. *Journal of Evolutionary Biology* 22.2, pp. 355–366. Link.

Moore, Richard (2017). “Convergent Minds: Ostension, Inference and Grice’s Third Clause”. *Interface Focus* 7.3, p. 20160107. Link.

Morton, Eugene S. and Richard G. Coss (2013). “Mitogenic Rays and the Information Metaphor: Transmitted Information Has Had Its Day”. *Animal Communication Theory*. Ed. by Ulrich Stegmann. New York: Cambridge University Press, pp. 207–231. Link.

Neander, Karen (2017). *A Mark of the Mental: In Defense of Informational Teleosemantics*. MIT Press.

Noel, Adam et al. (2017). "Using Game Theory for Real-Time Behavioral Dynamics in Microscopic Populations with Noisy Signaling". arXiv:1711.04870 [Physics, q-Bio]. arXiv: 1711.04870. Link.

O'Brien, Gerard and Jon Opie (2015). "Intentionality Lite or Analog Content?" *Philosophia* 43.3, pp. 723–729. Link.

Okasha, Samir (2018). *Agents and Goals in Evolution*. Oxford, New York: Oxford University Press.

Owren, Michael J., Drew Rendall, and Michael J. Ryan (2010). "Redefining Animal Signaling: Influence versus Information in Communication". *Biology & Philosophy* 25.5, pp. 755–780. Link.

Oxford English Dictionary (2020). *Oxford English Dictionary*. Link.

Pandey, Biswajit and Suman Sarkar (2017). "How Much a Galaxy Knows about Its Large-Scale Environment?: An Information Theoretic Perspective". *Monthly Notices of the Royal Astronomical Society: Letters* 467.1, pp. L6–L10. Link.

Parker, G. A. and J. Maynard Smith (1990). "Optimality Theory in Evolutionary Biology". *Nature* 348.6296, pp. 27–33. Link.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pfeifer, Jessica (2006). "The Use of Information Theory in Biology: Lessons from Social Insects". *Biological Theory* 1.3, pp. 317–330. Link.

Piccinini, Gualtiero and Andrea Scarantino (2011). "Information Processing, Computation, and Cognition". *Journal of Biological Physics* 37.1, pp. 1–38. Link.

Planer, Ronald J. and David Kalkman (2019). "Arbitrary Signals and Cognitive Complexity". *The British Journal for the Philosophy of Science*. Link.

Plath, Jenny A. et al. (2017). "Different Roles for Honey Bee Mushroom Bodies and Central Complex in Visual Learning of Colored Lights in an Aversive Conditioning Assay". *Frontiers in Behavioral Neuroscience* 11. Link.

Prabhakar, Arjun et al. (2017). "Dynamic Basis of Fidelity and Speed in Translation: Coordinated Multistep Mechanisms of Elongation and Termination". *Protein Science : A Publication of the Protein Society* 26.7, pp. 1352–1362. Link.

Preece, Kaitlyn and Madeleine Beekman (2014). "Honeybee Waggle Dance Error: Adaption [Sic] or Constraint? Unravelling the Complex Dance Language of Honeybees". *Animal Behaviour* 94, pp. 19–26. Link.

Price, Huw (2008). "Two Readings of Representationalism". Lecture. Descartes Lectures: Tilburg. Link.

Quastler, Henry, ed. (1953). *Essays on the Use of Information Theory in Biology*. Urbana: Univ. of Illinois Press.

Rabinowicz, Wlodek (2015). "From Values to Probabilities". *Synthese*, pp. 1–29. Link.

Ramsey, William M. (2007). *Representation Reconsidered*. Cambridge University Press.

Reading, Anthony (2011). *Meaningful Information*. Vol. 1. Springer Briefs in Biology. New York, NY: Springer New York. Link.

Reddy, Michael J. (1979). “The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language”. *Metaphor and Thought*. Ed. by Andrew Ortony. Cambridge, UK: Cambridge University Press, pp. 254–83.

Rendall, Drew and Michael J. Owren (2013). “Communication without Meaning or Information: Abandoning Language-Based and Informational Constructs in Animal Communication Theory”. *Animal Communication Theory*. Ed. by Ulrich Stegmann. New York: Cambridge University Press, pp. 151–188. Link.

Rendall, Drew, Michael J. Owren, and Michael J. Ryan (2009). “What Do Animal Signals Mean?” *Animal Behaviour* 78.2, pp. 233–240. Link.

Rescorla, Michael (2013). “Millikan on Honeybee Navigation and Communication”. *Millikan and Her Critics*. Ed. by Dan Ryder, Justine Kingsbury, and Kenneth Williford. John Wiley & Sons, pp. 87–102.

Reznikova, Zhanna (2017). *Studying Animal Languages Without Translation: An Insight from Ants*. [Online]: Springer. Link.

Riley, J. R. et al. (2005). “The Flight Paths of Honeybees Recruited by the Waggle Dance”. *Nature* 435.7039, pp. 205–207. Link.

Rohwer, Sievert (1982). “The Evolution of Reliable and Unreliable Badges of Fighting Ability”. *American Zoologist* 22.3, pp. 531–546. Link.

Romer, Heiner (1993). “Environmental and Biological Constraints for the Evolution of Long-Range Signalling and Hearing in Acoustic Insects”. *Philosophical Transactions: Biological Sciences* 340.1292, pp. 179–185. Link.

Ryan, Michael J. (2013). "The Importance of Integrative Biology to Sexual Selection and Communication". *Animal Communication Theory*. Ed. by Ulrich Stegmann. New York: Cambridge University Press, pp. 233–255. Link.

Ryder, Dan, Justine Kingsbury, and Kenneth Williford, eds. (2013). *Millikan and Her Critics*. John Wiley & Sons.

Sarkar, Sahotra (2013). "Information in Animal Communication: When and Why Does It Matter?" *Animal Communication Theory*. Ed. by Ulrich E. Stegmann. New York: Cambridge University Press, pp. 189–205. Link.

Savage, Leonard J. (1954). *The Foundations of Statistics*. Courier Corporation.

Scarantino, Andrea (2015). "Information as a Probabilistic Difference Maker". *Australasian Journal of Philosophy* 93.3, pp. 419–443. Link.

Schneider, Nicole A. and Michael Griesser (2013). "Incubating Females Use Dynamic Risk Assessment to Evaluate the Risk Posed by Different Predators". *Behavioral Ecology* 24.1, pp. 47–52. Link.

Schulte, Peter (2015). "Perceptual Representations: A Teleosemantic Answer to the Breadth-of-Application Problem". *Biology & Philosophy* 30.1, pp. 119–136. Link.

Schürch, Roger, Margaret J. Couvillon, and Madeleine Beekman, eds. (2016). *Ballroom Biology: Recent Insights into Honey Bee Waggle Dance Communications*. Frontiers Media SA. Link.

Schürch, Roger and Francis L. W. Ratnieks (2015). "The Spatial Information Content of the Honey Bee Waggle Dance". *Behavioral and Evolutionary Ecology* 3, p. 22. Link.

Scott-Phillips, T. C. (2008). "Defining Biological Communication". *Journal of Evolutionary Biology* 21.2, pp. 387–395. Link.

Searcy, William A. and Stephen Nowicki (2005). *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton University Press.

Seeley, Thomas D., Alexander S. Mikheyev, and Gary J. Pagano (2000). "Dancing Bees Tune Both Duration and Rate of Waggle-Run Production in Relation to Nectar-Source Profitability". *Journal of Comparative Physiology A* 186.9, pp. 813–819. Link.

Shannon, Claude E. (1948a). "A Mathematical Theory of Communication". *Bell System Technical Journal* 27.3-4, pp. 379–423, 623–656.

— (1948b). "A Mathematical Theory of Communication (Part 1)". *Bell System Technical Journal* 27.3, pp. 379–423. Link.

— (1948c). "A Mathematical Theory of Communication (Part 2)". *Bell System Technical Journal* 27.4, pp. 623–656. Link.

— (1959). "Coding Theorems for a Discrete Source with a Fidelity Criterion". *Collected Papers*. Wiley-IEEE Press, pp. 325–350.

Shannon, Claude E. and Warren Weaver (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press. Link.

Shea, Nicholas (2007). "Consumers Need Information: Supplementing Teleosemantics with an Input Condition". *Philosophy and Phenomenological Research* 75.2, pp. 404–435. Link.

— (2013). "Millikan's Isomorphism Requirement". *Millikan and Her Critics*. Ed. by Dan Ryder, Justine Kingsbury, and Kenneth Williford. John Wiley & Sons, pp. 63–86.

— (2018). *Representation in Cognitive Science*. Oxford University Press.

Shea, Nicholas, Peter Godfrey-Smith, and Rosa Cao (2017). "Content in Simple Signalling Systems". *The British Journal for the Philosophy of Science*, pp. 1–27. Link.

Sheehan, Michael J. and Thore J. Bergman (2016). "Is There an Evolutionary Trade-off between Quality Signaling and Social Recognition?" *Behavioral Ecology* 27.1, pp. 2–13. Link.

Sims, Chris R. (2016). "Rate-Distortion Theory and Human Perception". *Cognition* 152, pp. 181–198. Link.

Skyrms, Brian (2010). *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press.

Snyder, I. (1985). "Evolution of the Venus Fly Trap". *Carnivorous Plant Newsletter* 14.4, pp. 108–109.

Soare, Robert I. (1996). "Computability and Recursion". *The Bulletin of Symbolic Logic* 2.3, pp. 284–321. Link.

Sperber, Dan and Deirdre Wilson (1986). *Relevance: Communication and Cognition*. Second. Oxford; Cambridge, MA: Blackwell Publishers.

Stegmann, Ulrich E. (2004). "The Arbitrariness of the Genetic Code". *Biology and Philosophy* 19.2, pp. 205–222. Link.

— (2009). "A Consumer-Based Teleosemantics for Animal Signals". *Philosophy of Science* 76.5, pp. 864–875. Link.

— ed. (2013a). *Animal Communication Theory: Information and Influence*. New York: Cambridge University Press. Link.

— (2013b). "On the 'Transmission Sense of Information'". *Biology & Philosophy* 28.1, pp. 141–144. Link.

— (2015). "Prospects for Probabilistic Theories of Natural Information". *Erkenntnis* 80.4, pp. 869–893. Link.

Sterelny, Kim (1995). “Basic Minds”. *Philosophical Perspectives* 9, pp. 251–270. Link.

— (2003). *Thought in a Hostile World: The Evolution of Human Cognition*. Wiley.

— (2013). “Human Behavioral Ecology, Optimality, and Human Action”. *Evolution of Mind, Brain, and Culture*. Ed. by Gary Hatfield and Holly Pittman, pp. 303–324.

Sterelny, Kim and Ben Fraser (2017). “Evolution and Moral Realism”. *The British Journal for the Philosophy of Science* 68.4, pp. 981–1006. Link.

Tao, Terence (2013). *An Introduction to Measure Theory*. Link.

Templeton, Christopher N., Erick Greene, and Kate Davis (2005). “Allometry of Alarm Calls: Black-Capped Chickadees Encode Information About Predator Size”. *Science* 308.5730, pp. 1934–1937. Link.

Vanhooydonck, Bieke, Anthony Herrel, and Duncan J. Irschick (2006). “Out on a Limb: The Differential Effect of Substrate Diameter on Acceleration Capacity in Anolis Lizards”. *Journal of Experimental Biology* 209.22, pp. 4515–4523. Link.

von Frisch, K. (1948). “Gelöste Und Ungelöste Rätsel Der Bienensprache”. *Naturwissenschaften* 35, pp. 38–43. Link.

von Frisch, Karl (1950). *Bees: Their Vision, Chemical Senses, and Language*. Ithaca, NY: Cornell University Press.

— (1952). “Die Wechselseitigen Beziehungen Und Die Harmonie Im Bienenstaat”. Colloques Internationaux Du CNRS XXXIV. Paris, pp. 271–292.

von Neumann, John and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Science Editions, J. Wiley.

Wagner, E. O. (2012). “Deterministic Chaos and the Evolution of Meaning”. *The British Journal for the Philosophy of Science* 63.3, pp. 547–575. Link.

Walton, Kendall L. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Harvard University Press.

Wiener, Norbert (1948). *Cybernetics; or, Control and Communication in the Animal and the Machine*. Second. New York: M.I.T. Press.

Wiley, R. Haven (1983). "The Evolution of Communication: Information and Manipulation". *Animal Behaviour: Communication*. Ed. by Tim Halliday and Peter James Bramwell Slater. Vol. 2. Oxford: Blackwell Scientific Publications, pp. 156–189.

— (1994). "Errors, Exaggeration, and Deception in Animal Communication". *Behavioral Mechanisms in Evolutionary Ecology*. Ed. by Leslie Real. University of Chicago Press, pp. 157–189.

— (2013a). "Communication as a Transfer of Information: Measurement, Mechanism and Meaning". *Animal Communication Theory*. Ed. by Ulrich Stegmann. New York: Cambridge University Press, pp. 113–132. Link.

— (2013b). "Signal Detection, Noise, and the Evolution of Communication". *Animal Communication and Noise*. Ed. by Henrik Brumm. Animal Signals and Communication. Springer, Berlin, Heidelberg, pp. 7–30. Link.

— (2017). "How Noise Determines the Evolution of Communication". *Animal Behaviour* 124, pp. 307–313. Link.

Williams, Stephen E. (1976). "Comparative Sensory Physiology of the Droseraceae-The Evolution of a Plant Sensory System". *Proceedings of the American Philosophical Society* 120.3, pp. 187–204. Link.

Wilson, Edward O. (1962). "Chemical Communication among Workers of the Fire Ant Solenopsis Saevissima (Fr. Smith) 2. An Information Analysis of the Odour Trail". *Animal Behaviour* 10.1–2, pp. 148–158. Link.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. John Wiley & Sons.

Woodward, James (2003). *Making Things Happen*. Oxford University Press. Link.

— (2019). "Scientific Explanation". *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University. Link.

Woody, Andrea I. (2015). "Re-Orienting Discussions of Scientific Explanation: A Functional Perspective". *Studies in History and Philosophy of Science Part A* 52, pp. 79–87. Link.

Yockey, Hubert P. (1958). *Symposium on Information Theory in Biology, Gatlinburg, Tennessee, October 29-31, 1956*. Pergamon Press, Symposium Publications Division.

Young, R. and H. Bremer (1976). "Polypeptide-Chain-Elongation Rate in Escherichia Coli B/r as a Function of Growth Rate". *Biochemical Journal* 160.2, pp. 185–194. Link.