Intentionality and Neuroscience

Toward a Translation Between Mind and Brain State Descriptions

Dissertation der Graduate School of Systemic Neurosciences
der Ludwig-Maximilians-Universität München

Joachim Lipski

München, 2016

Diese Dissertation wurde angefertigt unter der Leitung von
Prof. Dr. Stephan Sellmaier, Prof. DDr. Hannes Leitgeb,
Prof. Dr.-Ing. Stefan Glasauer und PD Dr. Martin Rechenauer
im Bereich der Graduate School of Systemic Neurosciences
an der Ludwig-Maximilians-Universität München.

Erstgutachter: Prof. Dr. Stephan Sellmaier
Zweitgutachter: Prof. DDr. Hannes Leitgeb

Tag der Abgabe: 1. Dezember 2015

Tag der mündlichen Prüfung: 31. Mai 2016

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.
Die Dissertation wurde weder ganz noch teilweise bei einer anderen Prüfungskommission vorgelegt. Ich habe noch zu keinem früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

München, den

Table of Contents

List of Figures

Acknowledgements

I wish to extend my sincere thanks to:

My supervisors Stephan Sellmaier, Hannes Leitgeb, Stefan Glasauer and Martin Rechenauer for their valuable guidance and for challenging my preconceptions.

My friends and colleagues at the GSN and the Research Group for Neurophilosophy and Ethics of Neuroscience, especially Janett Triskiel, Lara Pourabdolrahim, David Kaufmann, Rey Francis Hernandez, Feli Selter, Steffen Steinert, Ondrej Havlicek, Ali Yousefi and Mario Günther.

Stephan Schleim, Michael von Grundherr, Florian Leiss, Maureen Sie and Jeanette Kennett for their helpful comments on early versions of this book.

The board and staff of the Graduate School of Systemic Neurosciences in Munich for providing an excellent work (and social!) environment, especially Benedikt Grothe, Lena Bittl, Maj-Catherine Botheroyd, Alexandra Stein, Julia Brenndörfer, Renate Herzog, Stefanie Bosse, Raluca Deac and Alex Kaiser.

The participants of the 2011 Summer School "Contemporary Psychology and the Moral Point of View" held by the Netherlands School for Research in Practical Philosophy, as well as of the 2014 GSN workshop "What is Neurophilosophy?", and of the GSN's 2011–2014 seminars at Venice International University, where I was able to present some of my early ideas for this book.

Jakob Steinbrenner, Gerhard Ernst, Ulrich Winko and Michael Zehetleitner for shaping my academic interests.

My parents and grandparents for their loving support.

Introduction

Representations are notorious troublemakers. Nevertheless, they and their close kin – information, function and semantic content – are prominently invoked in psychology, the neurosciences and the cognitive sciences in general. Representational states include our beliefs about the content of our fridges, our intentions about going to the movies, and our neural structures aimed at the recognition of faces. It is impossible to conceive of either of said fields as explaining what they do without invoking intentional terms: terms which involve being aimed at something or being about something. Uncovering the nature of intentionality and revealing its connection to research in cognitive neuroscience is what this book is about.

There is virtually no one in cognitive science today who would seriously contest that it is the brain in virtue of which we have mental states. Thus, descriptions of how the brain works play a crucial role in explaining intentional psychological properties. At its most basic, such descriptions are provided by neurobiology: by describing physicochemical properties of nervous systems. Yet, representational notions are not a proper part of the conceptual inventory of physics or any closely related science. Especially in virtue of their being tied to normativity and rationality, they go beyond what is describable physically or naturalistically. Consequently, it can seem puzzling how physical objects or structures come to be representational in the first place. This long-enduring puzzle has frequently been taken as an invitation to attempt to explain away representation as pertaining only to bogus properties (and the sciences which rely on them as pseudoscience), to attempt to naturalise representations (with mixed and far from uncontroversial results), or to mystify physics by endorsing panpsychism: by holding that mental properties are a basic feature of our physical universe. This book, however, endorses the view that the notion of representation yields significant explanatory value and should therefore be taken seriously, and that such an *intentional realism* ultimately needs to be reconcilable with a sober form of physicalism.

The most common way of characterising representations is by their role in symbolic systems. A symbol is composed of a material part (the "signifier") and its meaning (the "semantic content"). For example, some ink blots qualify as instantiating a specific word which in turn represents the word's meaning. This notion of representation is used as an

explanatory concept in intentional psychology. Intentional psychology uses attitudes (such as believing or desiring) toward semantic content in order to causally explain behaviour. For example, someone's desiring a beverage can cause their pouring a drink. Applications also extend into the animal realm, whenever we have reasons for believing animals to be capable of having certain mental states which can fulfill similar causally explanatory roles, or even into the realm of robotics or household applications (when we say that a robot goes left because it sees an obstacle on the right, or that a thermostat heats up the room because it believes it to be too cold).

While having such an intentional attitude need not itself involve using symbols or performing a symbolic action, I will argue that matters of symbolic ascription are in fact constitutive of content-ascriptions in intentional psychology. When psychological attitudes are ascribed to agents who are oblivious to symbolic ascriptive practice, I will call these states "sparse", but whenever they presume an agent's responsiveness to matters of rationality underlying such practice I will call them "rich". This distinction is meant to highlight that agents can act for reasons which they themselves are systematically oblivious to, and if this is so, then norms of rationality have no direct bearing on them. For example, if a thermostat does not act on the fact that the room is too cold, merely providing it with good reasons to heat it up will not sway it. In this case and many others, the ascribed form of content is sparse. However, when an agent's behaviour is shaped by these very norms, the corresponding content is richer, since it presupposes a different form of cognitive responsiveness. This distinction between rich and sparse content is integral to matters of ascribing content across different kinds of agents, and while intentional states can in an attenuated way be ascribed to thermostats, it is especially the rich kind of content-ascription which intentional psychology exploits.

I am also going to argue that the notion of an intentional mental state can only be understood against the backdrop of a psychological theory: mental representations are objects invoked to explain human behaviour, and so their meaning is determined by how any of them systematically explains it. Explanations in intentional psychology are causal and lawlike. So, understanding what an intentional mental state is rests on understanding what kind of lawlike inferences are supported by the ascription of such a mental state. Since parts of this argument require bolstering which may go beyond deeply rooted intuitions, I will also investigate alternative approaches – chiefly those holding that having or ascribing a mental state can be a kind of pretheoretical *brute fact* which we have immediate access to, such as by way of introspection – and go into their shortcomings.

The view that matters of meaning are interlocked with matters of intentional psychology in an intimate way has been popularly argued for in recent analytic philosophy. Advocates of this view developed arguments establishing a necessary connection between the abilities to have and ascribe intentional states, the knowledge of laws of intentional psychology and the mastery of matters of symbolic and mental meaning. To understand meaning it has been thought to be necessary to be able to ascribe mental states, and in order to ascribe mental states to others it should be necessary to know how to ascribe these states to oneself. Some of these interconnections immediately present themselves when considering examples: for anyone would obviously lack the competence to ascribe the belief that, say, grapes are sweet if they did not even know what kind of behaviour is typically explained by someone's having this belief in conjunction with a desire for eating sweet food. While an implication along these lines – that theoretical knowledge about the laws of intentional psychology is necessary in order to ascribe such psychological states – is relatively easy to swallow, it has also been argued that mental states cannot even be had without such competence. To make this point clear, I will partially rely on an argumentative strategy employed by the late Donald Davidson. The way he tied meaning itself to the ascription of intentional psychological states forms the bedrock for my claim that matters of meaning and matters of mental states are inseparable.

Analysing this intimate relationship between (both mental and non-mental) representation and intentional psychology is the subject of my first chapter. This relationship is not to be understood as just *any* kind of relationship which intentional psychology happens to maintain to other matters, but as essentially characterising the field. So, analysing this relationship is my method of choice for reconstructing what intentional psychology itself is. The chapter is structured thematically, as a logical introduction to the characteristics of intentional psychology, rather than by giving a one-by-one overview of important theories about the field. Nevertheless, this reconstruction will draw on many such theories from the contemporary and recent literature in order to clearly bring out its central issues.

In the cognitive sciences in general, representational concepts are used more loosely than in intentional psychology. On the one hand, intentional explanation informs all kinds of psychological theories: common categories of beliefs, desires, emotions, and so on, figure as singling out explananda in academic psychology. Yet, academic psychology is broader than intentional psychology, since its resulting explanations of such intentional phenomena go beyond the conceptual inventory of intentional psychology. That is, explanations of intentional states need not themselves invoke intentional states. For example, sitting at a dirty

desk has been found to influence moral judgments (Schnall et al. 2008a & 2008b), finding coins in a phone booth has been found to influence helping behaviour (Isen & Levin 1972), and the amount of meal breaks a judge has had before sentencing has been found to influence its severity (Danziger et al. 2011). Also common are explanations using so-called black-box models, schemas picking out causal factors of cognitive processes and characterising their interaction. For example, several types of marketing and environmental stimuli as well as buyer characteristics interact in models aiming to predict consumer behaviour. And when it comes to investigating processes of information transfer either in artificial systems or in actual neural networks, and to modelling how a specific output is or can be derived from a specific input, determining computational processes plays a central role. All of these approaches fall squarely into the methodology of the cognitive sciences, and all can be called "representational", while not necessarily invoking the notion of semantic content which intentional psychology invokes.

Since the cognitive sciences encompass both psychology and the neurosciences and are centrally taken to investigate matters of cognitive representation, one could expect that these differently employed forms of representation turn out to be offshoots of one *basic* notion of cognitive representation, and that the notions used in psychology and cognitive neuroscience are best understood in how they relate and contribute to such a basic notion. Regrettably this is not the case, since the claim that the cognitive sciences investigate forms of "cognitive representation" is not backed up by a basic definition of what is *cognitive* – a glaring omission Jesse Prinz calls "scandalous" (Prinz 2004: 41). So, what unites the various fields in the cognitive sciences is not that they all point toward a common notion of representation, but rather that they can all be (more or less loosely) treated as describing and explaining forms of processing of different kinds of mental representations. Some explain this processing in mathematical, some in logical, some in conceptual, and some in causal terms. That all of these forms of processing converge on their objects is, at least implicitly, due to the allusion to a common investigation of the mind. Again, there is no common definition of what the mind is, but the question "does this field of inquiry investigate the mind?" seems to yield a clearer answer than the question "does it investigate cognitive representation?" – if only because there is no intuitive notion of the latter, but a whole bunch of intuitive notions about the former. And the connections between all of these intuitive notions are made by the very sciences which are lumped together in virtue of these connections: Mental state descriptions explain behaviour, behaviour is a causal consequence of neuronal activity, this activity can be modelled in terms of computational processes, these computational processes pick out causal

factors of cognition, the models can be tested by behavioural psychology, and so on. So, while we can find these connections which illuminate the notion of cognition, we should not expect to find a shared notion of representation.

In the neurosciences, where information processing is conceived of in terms of electrical signals transduced in cellular networks, semantic content also has no direct explanatory role that could be considered as being akin to its role in intentional psychology. Yet, we can still find the term "representation" being used abundantly: Some neurons or neuronal networks are said to represent somatic states, others emotions, sights, sounds, smells, spatial locations, and so on. These differences in usage and explanatory role show that the term "representation" is far from being used uniformly across the diverse disciplines making up the cognitive sciences, and therefore, we shouldn't expect problems posed by adopting one notion to necessarily also be posed by adopting another. So, while neuroscience can explain phenomena of intentional psychology, it does not inherit the latter's form of representation as an explanatory concept.

Yet, despite some substantial differences we can find a common structure underlying the forms of representation in intentional psychology and those in cognitive and neuroscience which I am going to delineate. Crucially, all forms of representation require teleological principles which connect descriptions of intrinsic or organismic states in a functional way with their environment. In the second chapter I will investigate and characterise the role representations play in cognitive neuroscience and the different principles underlying them. Here we will also find out how the neuroscientific research of intentional states can be reconciled with the form of non-intentional explanation which neurobiology offers, and that this reconciliation suggests the possibility of localised translations between mental and neural descriptions. These translations are localised insofar as some of them may only apply to individual agents over a certain amount of time.

The notions of representation invoked by the two fields chiefly considered in this book, namely intentional psychology on the one hand and neuroscience on the other, form two extremes, insofar as one is used in semantic explanation, whereas the other is invoked for physical explanation. Investigating these should turn out to be instructive for other kinds of cognitive representations as well, since many of these can be placed into the same spectrum opened up by the two extremes. At the end of the first chapter I will briefly consider some of the psychological and cognitive states which are not the main focus of this book, but which can be related to and at least partially investigated and understood in terms of intentional states.

Overall, my aim is to give a comprehensive account of issues connected to the notion of representation and intentionality in cognitive science rather than attempting to solve any highly specialised problem. This approach is justified by the fact that intentionality, together with free will and consciousness, arguably poses one of the three biggest philosophical problems to current cognitive and neuroscience. And while matters of both consciousness and free will are generally seen as subjects which neuroscience might shed new light on but which aren't integral to the application of neuroscientific methods, matters of intentionality are in fact integral to the enterprise of cognitive neuroscience itself: cognitive research in general depends on the notion of representation. So, my broad approach reflects my express desire to reach beyond matters indigenous to philosophy. In writing this book I tried not to presume any specialised philosophical background knowledge on the part of the reader. However, since intentionality has been under extensive scrutiny for quite a while now, important issues revolving around it run the danger of seeming idiosyncratic and only explainable by appeal to philosophical tradition. Some of the topics treated in chapter I may suffer from this apparent idiosyncracy. However, I hope that each of its subsections will eventually make clear how each topic contributes to understanding intentionality. To facilitate selective reading, I am providing individual summaries at the end of each chapter. These densely retrace the points made in the individual subsections and are primarily meant to increase the book's accessibility by clearly marking where to find which argument or topic. A less dense recapitulation of the entire line of reasoning can be found in the conclusion.

As for the current state of the philosophical research of intentionality, much illuminating work has been done in 20th century analytic philosophy of mind, and this work informs many of the theoretical concepts invoked in the cognitive sciences today. However, the analytic philosophy of mental representation may be facing its twilight years. This is owed to the fact that many believe that all that can be said about this topic has already been said and consequently that much of what is currently said is a rehash: "Every conceivable position seems to have been occupied, along with some whose conceivability it is permissible to doubt. And every view that anyone has mooted, someone else has undertaken to refute. (…) But the chaotic appearances are actually misleading. A rather surprising amount of agreement has emerged, if not about who's winning, at least about how the game has to be played" (Fodor 1985: 76). So, while I won't attempt to reinvent the wheel in these pages, I do propose building a new chassis on top of a set of them. Or, to put it more bluntly: In light of a new quality of interaction between philosophers and empirical scientists, we can provide an

updated (while not radically revised) foundation for the notion of representation in cognitive science.

Before concluding this introduction, I wish to add a few words about the philosophical method and its integration into cognitive science. Joshua Knobe (2015) recently published some quantitative data on "what philosophers of mind actually do" (in fact, he not only considered work done in philosophy of mind, but also in epistemology, ethics and the philosophy of action). He diagnosed a shift from a less empirically informed philosophy to experimental philosophy and philosophy relying on empirical results. In comparing a sample of highly cited papers from 1960 to 1999 with a recent sample from 2009 to 2013, ratios change substantially (see Figure 1).

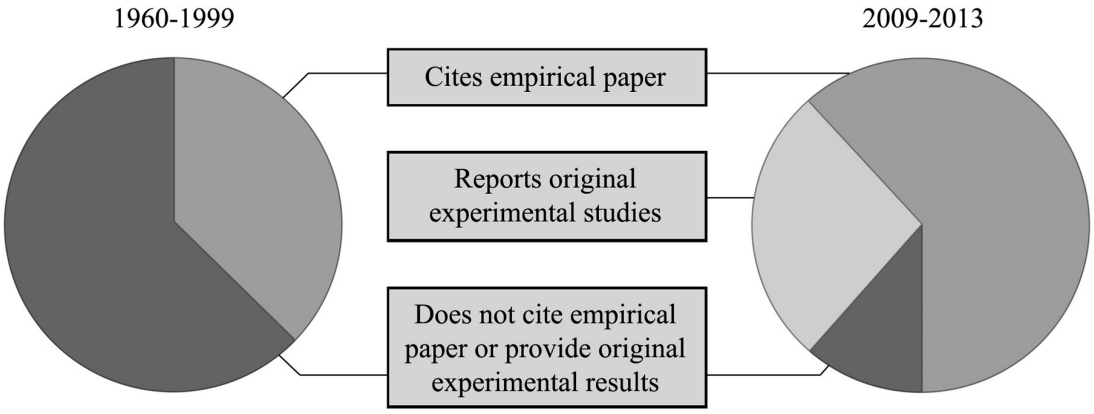

Figure 1: Distribution of philosophical methods in the late 20th and the early 21st century (after Knobe 2015).

One central tenet in recent neurophilosophy has been that (at least some) philosophical problems can be solved empirically. However, the basic idea underlying my investigation in this book is rather that empirical science always comes with a host of philosophical problems. So, while the methods of empirical scientists and the methods of philosophers differ, there is no question that their work can directly connect: It connects in working on common problems, and it connects insofar as much of the conceptual work done by philosophers is informed by the current state of science and vice versa. As Sellars has stated succinctly, "there has arisen the temptation (…) to confuse the sound idea that philosophy is not science with the mistaken idea that philosophy is independent of science" (Sellars 1997: 80, §39). But "what we call the scientific enterprise is the flowering of a dimension of discourse which already exists in what historicians call the "prescientific stage", and that failure to understand this type of discourse

"writ large" – in science – may lead, indeed, has often led to a failure to appreciate its role in "ordinary usage", and, as a result, to a failure to understand the full logic of even the most fundamental, the "simplest" empirical terms" (ibid.: 81, § 40).

So, we should neither indulge the view that our mental categories are insulated from the discoveries of cognitive science, nor jump to the conclusion that cognitive science pursues the goal of doing away with common mental categories. Being sympathetic to both a pragmatic scientific realism (i.e. that what is considered real depends on our scientific theories) as well as to intentional realism (i.e. the view that intentional mental states have genuine explanatory value), I advocate a moderate view regarding the relation of mental and neural categories: Namely that if we retain the teleological principles governing representational explanations, mental state descriptions could (empirically) turn out to be translatable to neural state descriptions and vice versa. As I'm going to show at the end of the second chapter, such a translation is feasible even if mental theories are not reducible to neural theories. Matters related to reductionism will pop up here and there, but generally, this book has been written with the hope that separating the question what role intentionality plays in cognitive and neuroscience from the question whether mental states can or should be reduced makes for a valuable perspective. While scientific insights might change our picture of the mind radically at some point in the future, we still need to work with what we have, and not with what we are promised.

I. Intentional Psychology

I.1. A Basic Characterisation

Intentional psychology is the practice of explaining phenomena by invoking mental states which have semantic content. It is the basis of our social interactions, colours the perception of our lives, lends purpose to our daily affairs, enables communication and kinship. We value friendly dispositions, expect sadness to be a consequence of a dear friend's passing, understand how someone in a desperate pinch can be moved to commit criminal deeds, and much more. All of this is because we can, and usually do, know what connections exist between objective events (such as someone's passing) and intentional mental states (such as their peer's sadness), and we know of the relation between such mental states and their potential effect or expression (such as subdued behaviour), which supplies observable evidence for the former's ascription. Indubitably, much of this knowledge is engrained deeply within ourselves and an essential part of what we think of as the human condition: we can hardly imagine human interaction without these familiar psychological laws, and are easily prompted to see them at work in our surroundings.

Consequently, we also trace many of these into the animal kingdom: We think dogs sad when their master neglects them, and we think monkeys angry when they receive cucumbers instead of the grapes they expected (Brosnan & DeWaal 2003) – and we deem their behaviour well justified in cases such as these. Conversely, we would find it bewildering if any sentient being wasn't bothered by being treated badly, or if anyone wouldn't respond with joy to being reunited with a loved one. There are exceptions, of course, but none that would disastrously undermine said relations' explanatory power in regard to rational behaviour. It is the theoretical integration of these and similar cognitive and behavioural states, events, processes and the systematic relations between them which I subsume under the monicker "intentional psychology".

With some exceptions which I will explore shortly, intentional psychology is usually taken to explain a wide variety of complex human behaviour in terms of the interaction of "propositional attitudes", i.e. attitudes such as beliefs, desires or intentions toward propositions. A proposition is what is expressed by a that-clause. For example, I may have any given attitude toward the proposition that *it rains* – I may believe, desire or, if I also believe myself to be a rainmaker, even intend it. If I do have one such attitude, such as the belief that it rains, then it does not make a difference to the truth of the mental state ascription *that I believe that it rains* whether it is ascribed in English or in French: thus, the propositional attitude is not simply to be identified with a linguistic entity, such as a sentence,

but with its meaning. This terminology and form of analysis harkens back to Bertrand Russell's writings: "What sort of name shall we give to verbs like 'believe' and 'wish' and so forth? I should be inclined to call them 'propositional verbs'. This is merely a suggested name for convenience, because they are verbs which have the *form* of relating an object to a proposition" (Russell 1918: 227). "[P]ropositional attitudes like belief, desire, intention; being pleased, astonished, afraid, or proud that something is the case; or knowing, remembering, noticing, or perceiving that something is the case" (Davidson 2001b: 3) have been a mainstay of 20th century analytic philosophy, and they continue to be a central object of analysis.[1]

As Robert Cummins concisely states, explanations of behaviour in terms of beliefs, desires and intentions ("BDI" for short) are

> "by far the most familiar explanatory model [in contemporary psychology]. It is the model of common sense psychological explanation, as well as Freudian psychodynamics, and a great deal of current developmental, social and cognitive psychology. It is what Dennett praises as explanation from the intentional stance, and what Churchland deplores as folk-psychology. (Dennett, 1987; Churchland, 1981.) Underlying BDI is a set of defining assumptions about how beliefs, desires and intentions interact. These assumptions are seldom if ever made explicit, just as one does not make explicit the mechanical assumptions about springs, levers and gears that ground structural explanations of a mechanical machine. Everyone knows that beliefs are available as premises in inference, that desires specify goals, and that intentions are adopted plans for achieving goals, so it doesn't have to [be] said explicitly (except by philosophers)" (Cummins 2000: 127).

In what follows, I will treat propositional attitudes, the central terms or objects of intentional psychology (or of what Cummins calls BDI-explanations), as a subset of intentional mental states. It is plausible to assume that not all mental states are intentional, and that not all intentional states are propositional attitudes. The former is owed to a descriptive inventory of the mental states we're capable of having: For example, Searle lists "a pain, ache, tickle, or itch" (Searle 1979: 74) as non-intentional mental states, and it seems prudent to say that these states are mental as well as non-intentional, insofar as they are "not 'about' anything, in the way that our beliefs, fears, etc. must in some sense be about something" (ibid.). The latter, however, is at least partly owed to a theoretical or terminological decision. Mental states which are said to be intentional but non-propositional are usually those whose expression cannot be related to the mental mode or attitude by way of a that-clause. Examples

[1] See e.g. Quine 1980: ch. 1, 1956, 1960: 200 ff., Sosa 1970, Kripke 1979 & 1980, Lewis 1981, Davidson 2001a: ch. 2 & 7, Fodor 1989: ch. 1, Fridland 2015.

are liking, seeing, loving, hating, smelling, and so on: emotions and experiential states. "Perhaps we should explain what it is to have such an experience in terms of a propositional-attitude representation; but it's not obvious why we should" (Crane 2013: 104). So, we might, perhaps with much effort and in a rather roundabout way, be able to reconstruct non-propositional intentional states in propositional terms rather than allowing there to be genuine non-propositional intentional states. For example, smelling ketchup is much the same as perceiving *that* there is a smell of ketchup. So, "[w]hat holds for the propositional attitudes ought, it seems, to be relevant to sensations" (Davidson 2001b: 3) as well as to "knowledge, memory, attention, and perception as directed to objects like people, streets, cities, comets, and other non-propositional entities" (ibid.).

Rather than going into reformulation attempts, what matters to me is whether the intentional content of non-propositional states can be analysed in a similar way as that of propositional ones. Propositional attitudes can be easier to analyse than non-propositional ones because we have an idea of how the symbols expressing such propositions get their meaning (see I.4 and I.7.4). The content of a propositional and that of a non-propositional state are sometimes determined and/or acquired analogously, in which case I will treat them as explanatorily similar, even without conducting any attempt at reformulation. However, they need not be analogous, such as in the case of the cognitive aspects of know-how or skills (cf. Fridland 2015). Because of the close connection between intentionality and the propositional form I will focus on a kind of analysis which works for propositional states in general, and then see whether and how it can be applied to some non-propositional states. In any case, I am not going to insist that there are only those intentional mental states which can be analysed analogously to propositional attitudes. Rather, I believe that those that *are* in fact analysed this way are interesting enough to warrant their own analysis, that they are widely used in explaining human behaviour, and that we can for this reason treat them as a paradigmatic form of mental intentionality.

In this chapter I am going to develop a view of intentional psychology which is guided by insights about the practical requirements of mental state ascriptions, about the form and function of psychological explanation, and about semantic facts as pertaining to mental intentionality. For example, the fact that mental state ascriptions have an intersubjective function, namely the prediction, explanation and normative control of behaviour, implies that its observable aspects play a central role in their characterisation. Case in point, my friends' knowing my intention of meeting with them tomorrow at 8pm allows them to predict that I'll be there at 8, explain why I'll be there at 8, and scold me if I'm not there at 8.

I.2. Intentionality

The first question to ask when dealing with intentional mental states is: What is intentionality? The modern version of this originally medieval scholastic concept was introduced into academic psychology by Franz Brentano in the late 19th century. Brentano was key to turning psychology into an academic enterprise to begin with, and his students included famous psychologists such as Sigmund Freud[2] just as well as famous philosophers like Edmund Husserl. My own understanding of intentionality has its roots in Brentano's concept, but relies less on Brentano's phenomenological account than a semantic account which is closer to research in recent analytic philosophy.[3] For this reason, I will barely go into phenomenological aspects, but rather introduce and explore intentionality's semantic underpinnings.

However, before elaborating on what intentionality means, I should briefly mention what it does *not* mean: in psychology and action theory, the term is sometimes used in connection with agency, or as a property of actions or agents, as if it were derived from the word "intended" (as in "did he just push me intentionally?"). However, in the sense relevant for this book, intentionality is not directly related to the concept of agency, and no terminological relation obtains with the common usage of "intended".[4] Rather, 13th century theologian and philosopher Thomas Aquinas originally derived it from the Latin word *intentio*, with its corresponding verb being *intendere*, meaning "to aim at" (see Aquinas 1272/1952).[5] Intentionality, in this sense, is connected to the notion of reference, with the referenced object being the "aim" of an intentional state. So we should clearly distinguish between referring to something and intending something.

While there *are* plausible connections between intentionality and agency, this should not obscure the distinct meanings of the terms. One such plausible connections is that actions

[2] For Brentano's impact on Freud see Smith 1999: 9-15. In his letter to Silberstein from March 5th 1875, Freud refers to Brentano as "a damned clever fellow, a genius in fact" (Boehlich 1988: 95).

[3] For a critical history of intentionality since Brentano which considers both analytic and continental traditions see MacDonald 2012.

[4] The exception being that an intentional object of a mental state could be referred to as its "intended object". It should be understood that this use of "intended" can still be distinguished from its use in phrases such as "I did not intend to spill your drink". In the former sense, "intended" can be substituted by "referenced" or similar semantic notions, whereas in the latter, it cannot. Curiously, the word "mean" can be used just as ambiguously as "intended": it can figure in pointing out semantic relationships ("*La lune* means the moon in French") as well as intended actions ("I did not mean to spill your drink" – although the semantic notion could be brought out more clearly, e.g. by saying "I did not mean *spilling your drink*", or using quotation marks: "I did not mean 'to spill your drink'"). If not explicitly mentioned otherwise, I am using the semantic understanding of intentionality, not the one related to agency.

[5] For contemporary interpretations of Aquinas' view on intentionality see Kenny 1984 and Brower & Brower-Toland 2008. For an examination of medieval views on the connections between intentionality, cognition and mental representation as well as their legacy in modern thought see Klima 2014.

require intentions to precede or cause them, and intentions are themselves intentional mental states. Assuming a combination of theories of action and theories of mind which hold that any action requires intentionality, and that having any intentional mental state requires being an agent, instances of agency will always coincide with instances of intentionality. However, no such specific theory or combination of theories is assumed or implied by my use of the term "intentionality", and neither is any direct relation between it and any concept of agency. As Dennett points out:

> "When discussing the [ascription of intentional states] (...), the word 'intention' means something broader than [introspectible mental events which precede actions] (…). It refers to states that have content. Beliefs, desires, and intentions are among the states that have content. To adopt the intentional stance towards a person - it's usually a person, but it could be towards a cat, or even a computer, playing chess - is to adopt the perspective that you're dealing with an agent who has beliefs and desires, and decides what to do, and what intentions to form, on the basis of a rational assessment of those beliefs and desires. It's the stance that dominates Game Theory. When, in the twentieth century, John von Neumann and Oskar Morgenstern invented the theory of games, they pointed out that game theory reflects something fundamental in strategy. Robinson Crusoe on a desert island doesn't need the intentional stance. If there's something in the environment that's like an agent – that you can treat as an agent – this changes the game. You have to start worrying about feedback loops. If you plan activities, you have to think: 'If I do this, this agent might think of doing that in response, and what would be my response to that?' Robinson Crusoe doesn't have to be sneaky and tiptoe around in his garden worrying about what the cabbages will do when they see him coming. But if you've got another agent there, you do" (Daniel Dennett in Edmonds & Warburton 2015: 129).

Consequently, and following its etymological root, I will construe intentionality as the property of referring to something. Brentano held that mental states are characterised by "aiming" at referenced objects and different modes of referring to these:

> "Every mental phenomenon is characterised by what the Scholastics of the Middle Ages call the intentional (or mental) inexistence[6] of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In

[6] Here, "inexistence" means "inclusion".

> presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on" (Brentano 1973: 88).

For our purposes, this is to say that the form of, say, ascribing a belief consists in specifying *that* it is a belief and in specifying *what* is believed. In "Marcia believes that the sun is shining", "believes" specifies the mode of the intentional state, while "the sun is shining" specifies its content. This is exactly the form of what is traditionally called a "propositional attitude": what I just referred to as the intentional mode is the attitude taken toward what I just called the content, which consists in a proposition.

Tracing the concept of intentionality back to Brentano's writings, we find him insisting that this property is held exclusively by mental states. That is because in his work intentionality is crucially invoked as a criterion to distinguish physical from mental states (ibid.). Combined with the claim from Brentano's quote above, this means that according to his view, *all* mental states are intentional and *only* mental states are intentional. As mentioned, I assume no such thing (although I do believe that there is a strong link between mental states and intentionality). The central reason for abandoning this assumption is that non-intentional states have been playing important explanatory roles in contemporary psychology (see Bechtel & Wright 2009). I will take a closer look at some of these, and their relation to intentional states, in section I.9.

I.3. Objects of Mental Reference

Since we have construed intentionality as the property of referring to something, a consequent question is what this "something" consists in. Do mental states refer to external objects or internal cognitive states? For example, perhaps the content of my belief that the next supermarket is around the corner is properly explained in terms of an internal map in my brain – if so, my belief gains its content in virtue of persisting in an internal representation. Even so, the content of said belief still seems to be an external state of affairs: namely, that there is a supermarket around the corner. To make the distinction between internal and external objects, it might help to think of internal ones as objects constituted or characterised by intrinsic properties of the intentional state's subject, whereas external objects are rather constituted or characterised by extrinsic properties, such as physical properties not belonging to the physical basis (if there is one) of a subject's mental state. I will come back to this

distinction and its relevance for matters of mental content in more detail in section I.8. For now, I hope the idea of a "physical basis" of a mental state is clear enough, even if you happen not to share the assumption that there is one. (By the way, my talk about a physical basis here would also allow for a necessary but not sufficient basis, i.e. the claim that some physical basis is necessary for a mental state to exist, but that the mental state consists in more than this physical basis, and that from the mere presence of the physical basis we cannot infer the instantiation or nature of a mental state.)

One view which relates intentional states to external goings-on is one which identifies social practices or conventions as a root of meaning, so let's briefly consider the in this sense externalist notion that the fact that a mental state refers to a specific object has to be traceable back to certain social practices. How about the following example: my belief that a soccer team is made up of eleven field players depends on a social convention, namely the rule that a soccer team is made up of eleven field players. But this does not suffice for our purposes, since even if there were no such rule, I could still believe that a soccer team is made up of eleven field players – I would merely be wrong about it! Still, my wrong belief would nonetheless have meaning, namely the content of the proposition *that a soccer team is made up of eleven field players*. What presently matters is that the content has meaning, and even if this fact turns out to depend on a social convention, it cannot merely be a conventional soccer rule. It has to be another kind of convention, namely the convention that the symbols used to express said belief mean the associated content.

Thus, we need to be a bit more sublime in our choice of examples: let's say that if I believe that quarks can have several distinct spins, the fact that my belief refers to anything depends on the establishment of quantum physics. The establishment of quantum physics entails that its terms have meaning (by way of definition, description, stipulation or what have you), and conversely, without quantum physics, its terms would mean nothing. That is, you could write down the word "quarks", but even if you got lucky in catching some meaning by it, it could not be the meaning endowed by quantum physics. Could a Neanderthal have thought about quarks? What could possibly have qualified any thought of his to be a thought about quarks? Obviously, it cannot be the Neanderthal's utterance of the word "quarks", since, even if he had been able to produce a sound which would have sounded like the utterance of "quarks" to our ears, no convention would have been established to the effect that any such utterance means anything. Thus, being able to think about anything plausibly requires the establishment of some social conventions about symbolic expression.

On the other hand, simply performing a symbolic action which conforms with conventional standards isn't enough either. If I happen to raise my hand during a voting process, but am actually unaware that a voting is taking place, then I am not actually voting (perhaps I am voting in a legally binding sense, but I certainly do not mean to do so). And if I mispronounce "Porto" as "Bordeaux", then I am not actually referring to Bordeaux. So, meaning something depends on internal *and* external matters: symbolic conventions and the intention to exploit them in order to express something (see I.4.4).

But even if social conventions are necessary in order to be able to have intentional states, this does not imply that social conventions are somehow themselves part of an intentional state's referential object. Rather, pointing to its being embedded in social conventions would be part of a good explanation of an individual's intentional capacities. Now, if there was a way of tying the explanatory power of psychological theories, which builds on social conventions, to the object of reference, we might be getting somewhere. I will come back to this question in sections I.6 and I.7; what should for now be clear is how basic considerations of matters of intentionality lead to this question.

Historically, some of the issues related to the question whether intentional states refer to mental or non-mental objects (or to representational or non-representational objects) were initially further developed by Brentano's student Twardowski (cf. his 1977). (And Twardowski's work strongly influenced Meinong and Husserl, in whose work intentionality played a central role.) In response to the work of Twardowski, Meinong and Husserl the issue was also discussed by Brentano himself (Brentano 1973: 385). Its implications are connected to matters of ontology as well as of theoretical parsimony. For example, if one adopts the view that intentional mental states are directed toward external objects, what needs clearing up is what mental states referring to fictitious or imaginary objects are then directed toward (cf. Thomasson 1999).[7] Clearly, one has a harder time locating unicorns in the external world than in anyone's mind. However, trying to circumvent such problems by allowing mental (i.e. immanent or internal) objects into one's ontology results in an inverse problem by duplicating many already existing objects – for any thought about, say, Rome would not be directed toward the actual city of Rome, but toward a mental representation of Rome. Thus, for any external object our thoughts and psychological attitudes can be directed toward, one additional internal object would have to be introduced, effectively doubling our ontological

[7] Note that this question also points to the dichotomy between how objects actually are and how they are given to us, and that it is related to the distinction between extensional contexts (i.e. contexts in which substitution of an expression with any expression with the same extension preserves truth value) and intensional contexts (in which said truth-value preservation is not the case) in philosophy of language (see Quine 1980: ch. 8). See also section I.6.1.

inventory. Today, we can trace this discussion to theories of *embodied or enacted cognition* or the *extended, embodied or embedded mind*, which holds that "perception is cheap, representation expensive" (Haugeland 1995: 219), intending to do away with a lot of the representational objects which previous theories of the mind have allowed into their ontology.[8]

Once intentional states are identified with propositional attitudes, the object of mental directedness is tied to the content of the respective proposition. Focussing on propositional content means straying from Brentano's main objective, which consisted in taking stock of mental objects by way of introspection. While introspection can yield evidence for ascribing propositional attitudes, the nature of an attitude's content is not directly or necessarily bound to matters of introspection, and wanting to explicate mental content in terms of objects that appear in introspection would be seriously misleading (see section I.7 for the details). For Brentano, however, the endeavor of categorising phenomenological, subjective objects which appear in introspection was indeed the very pinnacle of empirical psychology (see Brentano 1995: 4) – a viewpoint which was most harshly criticised by behaviourists (see section I.7.3). Today, behaviourism itself has itself long come under criticism, but the current cognitive sciences can (and do) much more readily accommodate behaviourist notions than Brentano-type phenomenology: "Although the paradigms of classical and operant conditioning and associative learning theory were rejected by those (...) who founded modern cognitive science, they found a good home in cognitive neurobiology where versions of the two learning paradigms are widely used to this day in conjunction with electrophysiology experiments that are used to induce LTP [i.e. long-term potentiation]" (Sullivan 2014: 56 f.).

Ultimately, it is the view that the intentional properties of cognitive states are best explicated as being based in their functionality and their intersubjective aspect which keeps matters of consciousness separated from our present matters of investigation: "Perhaps consciousness isn't essential to mind in the way that cognition is. This does not make the problem of consciousness go away, but it does make it, provisionally, someone else's problem" (Cummins 1991: 20). This is not to say that matters of consciousness cannot be subjects of scientific investigation or that there is no place for phenomenological research in cognitive science. There is no denying that our mental life comes with characteristic experiential qualities, that current approaches to investigating them are promising (cf. Gallagher 2003 & 2012), and that neural structures can be meaningfully invoked to investigate questions of subjectivity (cf. Qin, Duncan & Northoff 2013) and consciousness

[8] Beside Haugeland 1995, see e.g. Chalmers & Clark 1998, Noë 2004: 75-122, Clark 2008, Menary 2010, Shapiro 2010. Also compare section II.7.2.

(Overgaard 2015). What *should* be denied is that free-floating subjective qualities could be all there is to an intentional state, or that the phenomenological qualities of any such state could by themselves ground its functionality: "[A]ny putative conscious experience should be the experience of an agent. The thought here is that we cannot make sense of the image of free-floating experiences, of little isolated islets of experience that are not even potentially available as fodder for a creature's rational choices and considered actions" (Clark & Kiverstein in Block 2007: 502; also compare Evans 1982: 158). Plus, why should evolution have equipped us with something that is so radically private as to be functionally inert? Any remotely plausible answer to this question will have to mention that such phenomenological qualities do indeed have a value that goes beyond providing subjects with subjective experiences; some value that is likely describable in functional terms; and if that is the case, then we are already set on the road toward the theory of the mind which I am going to argue for in the following sections.

On top of all this, describing current phenomenological research as the realisation of Brentano's aim to base empirical psychology wholly on introspection would be a far stretch indeed.[9] In this sense, his general enterprise has no systematic bearing on what follows, and even my use of intentionality owes little more than terminological and historic tribute to his legacy. For him, intentionality really wasn't about the "content of a proposition", if the latter means being identified by anything beyond introspective phenomenology.

I.4. Semantics

I.4.1. Basic Issues

The second question to ask when dealing with intentional mental states is: What is semantic content? So, we need to get clear on semantics: the study of meaning. The term "semantic" is used interchangeably with "as pertaining to meaning", such as "semantic properties" being "properties pertaining to meaning", a "semantic theory" being a "theory of meaning", and so on. Meaning, representation, denotation, information, aboutness, and so on, are all semantic terms. The term "representation" commonly refers to entities that are used in a multitude of symbolic practices, especially when talking about representation used in

[9] For a discussion of the relevance of private subjective states for psychological explanation and ascription see sections I.7.2 and I.9.4.

relation to propositional attitudes. There are forms of non-symbolic representation, and I will investigate some of these (in section I.9 and chapter II), but for much the same reason as why I am treating propositional attitudes as a paradigmatic starting point of my analysis, I am treating symbolic representation as a starting point of analysing representation per se (see I.4.4 for why mental meaning requires investigating symbolic meaning).

Apart from describing how representation works, and how methods of representation achieve their purpose, a central task of semantics lies in clearing up what meaning itself actually is. There have been numerous attempts to explain the meaning of "meaning". For example, to roughly sketch some influential theories (pertaining to potentially different forms of representation):

- Knowing the meaning of X means having a mental idea of X (Classic Internalism)
- The meaning of X is its use in a symbolic system (Wittgensteinian pragmatism, cf. his 1953: §43)
- The meaning of X are the conditions of its verification or the way X contributes to the verification conditions of a statement it figures in (Verificationism)
- Knowing the meaning of X means having a truth theory of the language in which X is an expression and knowing how X contributes to the truth of statements it figures in (Davidsonianism, see II.8.4.2)
- and so on.

This cursory classification is merely supplied for illustrative purposes. Presently, we neither need to subscribe to any of these views, nor distill one we can get behind. While the following sections will partially rely on claims about the nature of meaning, I am not going to make a dedicated effort to arrive at a unified account of what meaning is. What I am centrally going to argue for is that symbolic meaning requires conventional rules, that the use of symbols must be practical and fundamentally learnable, and that meaning is distinct from semblance and causation.

I.4.2. Semantic Mechanics

Symbolic representations are characterised by a reference-relation holding between appropriate kinds of objects: on the one end of the relation stands a *signifier* – the material

part of a *symbol* – and on the other its *meaning* (or its "semantic content")[10] That is, a signifier represents, is about, or refers to its *meaning*. For example, in case you happen to read this text on paper, then your understanding of what I am trying to tell you will depend on your perceiving a significant contrast between the parts of the paper on which shapes have been printed which are interpretable as letters (and ultimately words and sentences) and the "empty" non-letter space that surrounds them. In case you are reading it digitally, then contrasts between groups of pixels will do this job. In either case, the relevant contrast has been so designed as to fall under a certain representational scheme (i.e. letters, words and sentences) which can be reliably interpreted. What you take these to be standing for could be real objects, such as my writing the word "chair" could stand for any actual chair, or they could stand for abstract contents, such as scientific theories, or your ideas about any of these objects. In this sense, "meaning" can be ambiguous: it can stand for real objects or mental ideas (as reflected in the distinction between "reference" and "sense", cf. Frege 1892).

A *symbol* is what unites a certain group of material objects (i.e. those that qualify as a signifier) and specifies what the signifier refers to. So, both signifier and meaning are not actually singular objects, but should rather be treated as the specification of criteria which could be satisfied by an arbitrary amount of objects. For example, a certain structure of blots of ink, which is defined by its spatial and visual qualities, qualifies as forming a specific word which in turn represents this word's meaning. So, an indefinite amount of blots of ink can bear the characteristics necessary for being signifiers, and signifiers can in turn refer to, denote or represent any number of objects, just as a potentially infinite amount of instances of "chair"-tokens can refer to a potentially infinite amount of chairs.

There are, of course, forms of symbolic representations beyond linguistic ones such as letters, words and sentences. Some other common forms are sociopolitical representations (such as the way each US governor represents their state, or the way I could represent the GSN-LMU at a conference; cf. Pitkin 1967) and pictorial representations (such as "the *Mona Lisa* represents Lisa Gherardini"). In these cases, what makes a signifier a signifier may differ dramatically (having a certain similarity, being elected, having a certain causal history), but in general, the relational structure of the symbol applies. Characteristic for symbolic representation is also a certain conventionality and a pragmatic dimension. That is, which signifier represents which meaning is established conventionally, and what makes a representation a representation is it is used as such. This is why Charles W. Morris characterised a symbol as having three types of relations, namely *semantics* (the relation to

[10] Here I am adopting Ferdinand de Saussure's terminology (Saussure 1983: 67).

objects), *pragmatics* (the relations to people) and *syntactics* (the relation to other symbols; cf. Morris 1938), and why Charles Sanders Peirce defined "a sign as anything which is so determined by something else, called its object, and so determines an effect upon a person, which effect I call its interpretant, that the latter is thereby mediately determined by the former" (Peirce 1998: 478).

Pragmatics and conventionality are important in highlighting that anything can be used as a representation for anything, and that commonality of properties (i.e. similarity) between signifier and meaning is not sufficient for establishing a representation relation (cf. Goodman 1972: 437 – 446, Crane 2001: 348).[11] Of course, properties *can* be shared: Similarities between a picture and what is depicted can be exploited in order to establish the relation between the two. However, pictures can be completely dissimilar to what they depict (just as a person who represents Germany needs not resemble Germany, and nor does the word "Germany"). Whether the signifier is expected to share any property with the represented object also depends on convention. Scott McCloud has created an illuminating schematic for the kind of choices visual artists and designers face: Visual representations can be fashioned so as to evoke a certain (pseudo-)realistic semblance of the depicted objects, but they can just as well go for iconic abstraction, or for evoking properties of the artistic medium (see Figure 2). Crucially, similarity or a sharing of properties between two objects is not sufficient for one to be a representation of another. Generally, symbolic representations are representations due to their being artefacts fashioned and used for a representational purpose – and there are numerous methods for achieving this purpose.

No dichotomy or dualism is necessarily implied by a relational analysis of meaning alone. (Even when setting aside self-reference, a relationship between two identical objects, or instances in which symbols refer to other symbols.) In fact, when analysing meaning as an A referring to a B, A and B can still be similar in terms of belonging to the same kind, type or class of objects. For example, proponents of physicalism would probably want to have meaning understood as something "material" also, so calling the signifier the "material part of a symbol" would be misleading. And the option of incorporating meaning itself into a monist worldview like physicalism is certainly an option that's still on the table (depending on the strength of the physicalism endorsed, it can be conceived as a basic criterion for accepting a semantic theory). In any case, the mere difference between A and B implies nothing more substantial about the nature of A and B than their not playing the same semantic role.

[11] For problems surrounding notions of similarity regarding mental representation see Cummins 1991, ch. 3.

For this illustration, please refer to the print version or Scott McCloud's website (http://scottmccloud.com/4-inventions/triangle/index.html).

Figure 2: Scott McCloud's "Big Triangle".[12]

[12] For more information see http://scottmccloud.com/4-inventions/triangle/index.html.

As I've mentioned, in order for A to refer to B, A needs to fulfill specific criteria for counting as a signifier, so it needs to be somehow suited to refer to B. But it should be noted that the property of "being suited to refer to B" does not imply any inherent or immanent property on behalf of the signifier itself. In most paradigmatic cases of reference relations, signifiers signify merely by virtue of being used as signifiers, that is, by virtue of there being a convention about A's being used to refer to B – standard cases being words and images. No immanent or inherent property of any ink blot determines its being used as a letter; rather, once it is established that it can be used as a letter, then some of its intrinsic properties (such as its shape) qualify it as being interpretable as a letter.

Going into linguistic examples, words can refer to real, imagined or fictitious objects, to events or states-of-affairs, be they actual or counterfactual, and to many more things, while being wholly dissimilar to all of these, in structure as well as in appearance. In fact, all it takes for something to be referred to by a word is to establish the reference relation by way of convention. One instance of establishing such a relation can be a definition, i.e. introducing a word with the explicit purpose of referring to whatever one wants to have it refer to. Again, definitions are but one way of establishing such a relation; explicit definitions are not necessary for words to refer, and in many cases in which we use words to refer to things, we do so without ever having come across an explicit definition. We often learn to use words implicitly, by way of making guesses about how those around us use these words, and we collect various cues in the environment to support these guesses. And even when explicitly asking our peers about how they use their words, we should not expect their explanations to always be akin to definitions.

However, establishing reference relations is not a magic trick one can perform by merely conjoining word and object in the absence of any other conditions. Going back to our example of definitions, they usually do the trick because they are made in the context of an established convention, namely the convention that defining a newly introduced term (the *definiendum*) makes it refer to the content of the definition (the *definiens*), and they are made against the backdrop of a community whose common practice provides the fulfillment of necessary contextual conditions. For instance, if I were to define a new term right away, anyone able to understand this text will likely be predisposed to understand my definition, at least in principle. They may criticise my definition for being awkward, unnecessary or unwieldy, but even so, they will understand that by the definiendum I mean to refer to the definiens.

The specific formal and practical requirements for the establishment and the use of reference relations are embedded in our social practice of using symbols, and they have to be principally relatable and learnable in order for us to rely on them; but apart from these sketchy remarks, I will not pursue them any further here. What matters for now is that acknowleding that many reference relations primarily depend on conventions does not make the conjoining of signifier and signified a random matter, for the conventions may themselves rely on complicated and, to those exploiting the reference relationships, often opaque matters.

Relating reference to convention sets the former apart from relations such as semblance or causation. While many images are similar to what they refer to, and while many indicators are actually caused by what they indicate (such as a functioning speedometer's current display), both relations are neither necessary nor sufficient for establishing a reference relation. That is, we should not rule out that similarity and causality can play roles in picking what refers to what, or that reference relationships will at times exploit either, but we would be mistaken in expecting reference tout court to consist in nothing but similarity or causality: "For a representation is a representation of Pegasus not because it necessarily looks like Pegasus, nor because it is caused by Pegasus, but because it can be used to express thoughts (intentional states or acts) about Pegasus" (Crane 2001: 348).

Consider the case of the speedometer: if functioning correctly, the speedometer works because there is a causal link between display and speed. However, this is only because we have fashioned the speedometer in this way – the causal connection is not arbitrary, but rather, through technical ingenuity, the output of the speedometer is always interpretable as a symbol indicating the current speed. Thus, it would not do justice to analyse the speedometer only in descriptive terms of the speed causing a certain display. Two further things are required, namely that the speedometer has been fashioned to display speed (i.e. a norm) and that there is a symbolic convention which makes the speedometer's output interpretable as the display of a specific speed. Without being justified in assuming that the two latter requirements are met, we would have no reason to assume that the speedometer is displaying anything. If there was no symbolic convention to make the output interpretable, then all the causal relations in the world would not suffice. And if there was no norm for the speedometer to indicate the current speed, it would cease to be a speedometer. This is why clouds resembling the face of Zeus do not refer to Zeus, and why the patterns which male pufferfish create to attract females are not art (see section 2.4). They are not referential until they're made part of a symbolic practice which endows them with meaning (and sometimes, all it takes is someone saying: "look, Zeus!"). But, crucially, if nothing else can be said about a phenomenon except that it

instantiates a causal relation, or bears some semblance to an object, then we have no reason to assume it is one of reference.

I.4.3. Semantic Externalism and Causality

In spite of what I've just said about causality, I need to add a complication here, which touches the nature of how many symbols achieve their meaningfulness. For one form of *semantic externalism* consists in the claim that symbols refer to their meaning because the former stand in a causal relationship to the latter. For example, according to this view, the word "tiger" refers to tigers only because anyone's use of this word stands in an (occasionally long-winded) causal chain with someone's direct experience with a tiger – an event which caused that very animal to be named "tiger", which then causes any of us to refer to tigers by uttering the word "tiger" (the causal chain's length being of no immediate import).[13]

Now, this notion of reference as depending on causality does not conflict in any way with that of its depending on conventionality and the idea that reference is not readily reducible to causality. For, firstly, it is the initial "baptism" (cf. Kripke 1980: 96 ff.) which introduced conventionality: whoever encountered tigers first could have named them anything, and the community which adopted the use of the word could as well have rejected it.[14] And, secondly, even if causality plays a role in the way semantic externalism says it does, anyone's use of the word "tiger" does not indicate tigers the same way a speedometer indicates speed. While speedometres are also subject to conventionality – their display could take *any* readable shape or form –, their meaningfulness at any single point in time depends on an active, working and direct causal link between display (signifier) and speed (meaning). No such causal link is systematically assumed when talking about tigers. Of course, we could imagine a direct causal link such as this, namely when we are talking about a specific tiger which is currently present – i.e. indexing it THIS tiger HERE and NOW –, and we are updating our assertions about THIS tiger directly. In such cases, the truth of our assertions will depend on such an active and direct causal link. However, it is not reference itself which depends on this causal link; and since most of us will never be in such a situation, or may

[13] My points here are essentially an amalgamation of points made prominently in Kripke 1980 and Putnam 1981.
[14] Here, I have made two simplifications: namely, that the "baptism" was close enough to what can be described as a straightforward singular event, and that linguistic communities consequently adopt or reject the result of the naming. In reality, things will of course not be that easy. Yet, this does not touch their being conventional.

have never even been in the presence of a tiger beyond the confines of a zoo, neither does the meaningfulness of our assertions about tigers.

Commonly, reference is an asymmetric relationship, as are causal relationships. The word "tiger" refers to tigers, but usually, tigers don't refer to words. (I say *usually* because, since reference relationships can be introduced by way of convention, it might be possible to introduce the convention that real tigers are used as symbols for words. Absurd as it sounds, it is possible.) If causality plays a role in establishing meaning, then causal and referential relations should point in opposite directions: The symbol refers to the object, whereas the object causes the functionality of the symbol. Consider that in a Kripkean baptism, the tiger (plus some contextual conditions) causes the symbol "tiger" to refer to tigers . Thus, "tiger" refers to a tiger, whereas, going in the opposite direction, the tiger causes the semantic functionality (the "meaningfulness") of the symbol "tiger". In this sense, the two asymmetries are not *exactly* mirrored, since, if we take "A" to mean the symbol (the word "tiger") and "B" the referent (the actual tiger), then A means B, but B did not simply cause A. Rather, B caused *that A means B*. However, the direction is still reversed, since it is less likely that the symbol causes its referent. This much seems clear in the case of tigers and tiger-symbols, since no tiger-symbol has ever literally caused a tiger. (What does it mean to "cause a tiger"? Whatever it means, no symbol seems to have accomplished that.)

However, even in the example just invoked, we did not require that the tiger alone cause the meaningfulness of the word "tiger", but rather, the tiger plus some contextual conditions. If we accept that, then we can accept that the symbol "tiger" could cause the tiger plus some contextual conditions; for instance, the use of the symbol in a tourist's exclamation "let's go look for a tiger!" can cause the tiger to be present in their vicinity (where the tiger's presence are the contextual conditions, much as they constitute the contextual conditions in a "baptism"). In fact, we should suppose imperatives and exclamations to usually work this way: If we use a specific symbol in either, we should expect to bring about the referent of this very symbol (plus contextual conditions). If that wouldn't work, we would not be motivated to use imperatives at all. The exclamation "let's cause a riot!" is supposed to bring about a riot, after all.[15] Crucially, though, the fact that imperatives using certain symbols bring about their referents does nothing for the reference relation itself, since, logically, the reference relation has to be presupposed for an imperative to bring about the used symbols' referents. If the symbol "riot" would not refer to riots, then shouting "let's cause a riot" could not cause a

[15] This may also work in other linguistic forms besides imperatives and exclamations. Imagine a culture in which mentioning riots causes riots because of social conventions (perhaps mentioning riots is a taboo, and breaking taboos causes riots). In this case, mere mention of a symbol causes its referent.

riot (at least not in virtue of the exclamation's semantic properties). Therefore, the shouting causing the riot is not the basis of the reference relation, but vice versa. The fact that these sorts of causes are, in this sense, *semantically inert* establishes the reversed asymmetry I was aiming to get at: If causes play a role in semantics at all (at least in Putnam's and Kripke's sense), then they should go in one direction, namely from the referent to the reference relation (from the tiger to the symbol "tiger"), and not vice versa. Whereas the symbol refers to its referent, and not vice versa: Tigers do not refer to the word "tiger", but rather, the word "tiger" refers to a tiger. (Self-reference constitutes an exception, since "A → A" also implies "A ↔ A".)

I.4.4. Mentalism

So far, in probing the phenomenon of reference, I have been talking about symbolic reference only. But what bearing does symbolic reference have on intentionality, or more specifically, on the mind? One reason is straightforward: If intentionality is a property of certain mental states, and intentionality is properly analysed in terms of symbolic reference, then shedding light on symbolic reference is a proper part of analysing the respective mental states. However, I have not said enough about why intentionality should be properly analysed in terms of *symbolic* reference. Perhaps mental reference is completely different from symbolic reference, and if that is the case, then we could just ignore the latter.

But before I go into that, I should also mention that a second thread can be construed in the reverse direction: not from mental to symbolic reference, but from symbolic to mental reference. If none of its inherent qualities suffice for qualifying a physical object to function as a signifier, one question seems pressing: If we use them as such, how exactly do they get imbued with meaning? How do regular objects get transformed into symbols? What turns ink blots on paper or pixel arrays on screens into letters, what makes Picasso's "Guernica" refer to the horrors of war, what makes flags stand for countries, and what can turn a catastrophe's survivor into a symbol of hope? One answer to these questions is: The mind. And one way to interpret this answer is to hold that symbolic meaning is derived from mental processes: that they *only* have meaning insofar as they are used by people capable of intentional thoughts, and insofar as they are used in order to express and communicate these.[16] This comparably weak point is already implied by accepting the pragmatist view that there are no symbols

[16] See Chisholm 1958, Haugeland 1981, Searle 1980, 1983, 1992, Fodor 1989, Cummins 1991: 21 ff.

without their being used. For instance, this view implies that if there was no one to use words as signifiers, then words could not mean anything (or rather, they would not exist as such). If 80 million years ago, winds had formed shapes that resembled the letters B-E-N-C-H out of sand on a distant planet, they could not possibly have referred to any bench – in the absence of someone to use them as a signifier, and in the absence of a symbolic system which enables anyone to use them as such, they could not refer at all.

However, we can also opt for a considerably stronger view and not merely hold that something's being a symbol implies its being used in a symbolic practice by cognisers, but, more crucially, that symbolic representation is *derived* from mental intentionality. According to this view, mental intentionality is the primary form of intentionality, whereas symbolic intentionality is secondary. John Searle takes one such route when distinguishing between *intrinsic intentionality*, which is the intentionality exhibited by an individual's mental processes, and the intentionality exhibited by symbols, which is derived from these mental processes. To illustrate this point, he says that the French statement "J'ai grand faim en ce moment" is "derived from the intrinsic intentionality of French speakers. That very sentence might have been used by the French to mean something else, or it might have meant nothing at all, and in that sense its meaning is not intrinsic to the sentence but is derived from agents who have intrinsic intentionality [i.e. French speakers who exhibit the intentional desire of hunger]. All linguistic meaning is derived intentionality" (Searle 2000: 93). Similarly, Cussins points out that

> "[t]he theory of content – in terms of which we explain what content is – locates the notion with respect to our notions of experience, thought, and the world. But it is important to see that this is consistent with the notion of content being applied (though not explained in terms of) states which are not states of an experiencing subject. There are derivative uses of the notion in application to the communicative products of cognition, such as speech, writing, and other sign-systems, or to non-conscious states of persons such as subpersonal information-processing states, but these uses must ultimately be explained in terms of a theory of the primary application of content in cognitive experience" (Cussins 2003: 133).

Thus, the two opposing views we can take at this intersection coincide with the difference of explanatory direction: Do we explain symbolic intentionality by mental intentionality or vice versa? Searle, Cussins, Fodor and many others take the former route. Their theories, which take the foundation of meaning to be found in the mind, are consequently called mentalistic. One reason for adopting such a theory is holding the classic

Gricean view that the use of symbols is essentially pragmatic and therefore dependent on intentions. For example, by uttering "you're obstructing the view" toward a person sitting in front of me at a cinema, I do not mean to make a simple statement as evidenced by the literal meaning of the uttered sentence; rather, the intended meaning is to get this person to clear the view (compare Grice 1989: 86-116 & 213-223). In this sense, intentions are clearly indispensable when it comes to properly understanding the meaning of utterances (and other forms of symbolic expression). Thus, on this view, symbolic meaning is explained in reference to mental states. Another reason is a view like Fodor's: the view that the basis of intentional mental states are quasi-linguistic (i.e. word- or sentence-like) structures in the mind. This view has appropriately been dubbed the "Language of Thought" hypothesis (or LOT for short; cf. Fodor 1975)[17]. If this claim is true, then it is plausible to believe that such mental entities constitute the basis for non-mental sentences (like those in a book) as well.

Now, first off, I agree that pointing toward mental intentionality is necessary to get clear on non-mental intentionality. This is because symbols do require cognitive operations in order to mean something, operations whose investigation falls squarely into the psychological domain. There certainly are other (sociological and linguistic) dimensions to the question how symbols are imbued with meaning, but what psychology can contribute is to identify and characterise the bases for our ability to interpret symbols in the form of cognitive processes or mechanisms (such as those necessary for reading etc.). The skills employing such mechanisms are taught and honed intersubjectively, requiring established conventions, which is why analyses of the social contexts, and the role they are playing, are needed, and why individualistic analyses of such cognitive competences of individual interpreters are ultimately not sufficient for a full understanding of their nature (see also section I.8).

Although intentionality depends on such mental operations, these are themselves non-intentional: being apt at reading, i.e. cognitively transforming blots or pixels into letters, is not itself an intentional state. As Searle has pointed out, such skills belong to a background of "nonrepresentational mental capacities that enable all representing to take place" (Searle 1983: 143):

> "Think of what is necessary, what must be the case, in order that I can now form the intention to go to the refrigerator and get a bottle of cold beer to drink. The biological and cultural resources that I must bring to bear on this task, even to form the intention to perform the task, are (considered in a certain light) truly staggering. But without these resources I could not form the intention at all: standing, walking, opening and closing doors, manipulating bottles,

[17] Also see his 1970, 1989, 1994, 1998, 2008.

> glass, refrigerators, opening, pouring and drinking. The activation of these capacities would normally involve presentations and representations, e.g., I have to see the door in order to open the door, but the ability to recognise the door and the ability to open the door are not themselves further representations" (ibid.; also see Radman 2012).

When internalised, such abilities are skills of the same sort as those we employ when using tools: Once we are apt at driving a car, we can do so without consciously thinking about, say, how to change gears.[18] Rather, apt drivers have internalised enough facts about controlling their car in order to be able to *just do* it. In difficult and/or novel situations, which require their conscious attention (and which afford them enough time), drivers will switch from merely reacting to reflecting (i.e. to conscious/deliberate/cognitive control). Analogously, whenever readers try to decipher a bad printout, they will switch from merely recognising letters to actively trying to interpret them (by any number of strategies, many of them requiring conscious reflection). And we all know what trying to read while being distracted or tired is like. Sometimes, we fail at yielding a semantic output, and we will just have to read it *again*. So, while interpreting symbols requires cognitively outputting semantic representations, these will still be based on automatic and unconscious non-intentional mental skills.

These general remarks are, I believe, sufficient to establish a general agreement with Searle; and in no way do I believe that Grice is wrong in stressing that literal meaning is usually not sufficient to grasp intended meaning, and that thus, pragmatics are essential to intentionality. However, agreeing with these points still doesn't imply mental intentionality's primacy. Note that the mental states underlying non-mental intentionality in the two examples invoked at the outset should properly be construed as intentions for A to mean B, where A is an utterance (i.e. a non-mental symbol) and B its intended (mental) meaning. In Searle's case, A stands for "J'ai grand faim en ce moment", whereas in the Gricean example, it stands for "you're obstructing the view". At a first glance, it sounds as if what Searle had in mind was claiming that the mental state of being hungry was meant to explain the utterance "J'ai grand faim en ce moment", and that what Grice points to is that the desire to get someone to get out of the way explains the utterance "you're obstructing the view". But this is only part of the story. If we were to describe the explanatory states fully, we would need to add further intentions, namely "intending the sentence 'J'ai grand faim en ce moment' to imply being

[18] For a more elaborate distinction about the form(s) of consciousness involved here, see Armstrong 1981 (in particular his example of the long-distance truck driver). My notion of consciousness here does not encompass, say, perceptual consciousness in Armstrong's sense, but requires consciously entertaining intentions.

very hungry at the moment" and "intending the utterance 'you're obstructing the view' to imply that the person obstructing the view should get out of the way". Since these intentions are directed towards non-mental symbols (namely the utterances), we can see that in both cases the non-mental symbol is constitutive of having either mental intention. If there were no symbolic representation, no one could have a mental intention that was *about* it (i.e. we could not intend A to mean B if there were no As). Thus, there can be no understanding the mental intention without understanding the nature of symbols; and thus, mental intention cannot in any substantial way be primary to symbolic intentionality.

To be clear on this point: In Searle-like cases, the mental intention that was supposed to be primary to the intentionality of the sentence "J'ai grand faim en ce moment" is "I am very hungry right now" – a mental state that does not presuppose symbolic intentionality. The problem is, the notion of mentalistic primacy was construed as explanatory primacy, and as holding that the mental intentionality explains the sentential intentionality. However, the mental state of being hungry does not by itself explain why "J'ai grand faim en ce moment" refers to the speaker's being very hungry (and that was what the deference to the intentional mental state was supposed to achieve). In fact, it only does so if the speaker *also* intends to use the sentence as meaning her current hungriness. Of course, the mental state of hunger explains the utterance causally; that is, a Frenchman's hunger causally explains his utterance "J'ai grand faim en ce moment". But this is just to say that expressing one's hunger is, after all, an action that is to be explained, other things being equal, by the mental state of being hungry. However, this was not the point. The point was to explain how "J'ai grand faim en ce moment" gets its meaning, not why it is sometimes uttered by Frenchmen. And as pointed out, the full explanation would have to add that the Frenchman intended to express his hunger with a symbolic act, namely uttering "J'ai grand faim en ce moment". So, strictly speaking, there's not just one cognitive state explaining the utterance, namely the hunger, but also another cognitive state: that of knowing that one can express being hungry symbolically by uttering "J'ai grand faim en ce moment". Without this knowledge, the mental state of hunger alone could not explain the utterance.

Thus, it is true that we should point toward cognition when seeking to explain symbolic intentionality; but in cases as the ones just discussed, we also need to point toward symbolic intentionality when explaining cognition. This is because being apt at symbolic actions such as speaking a language, or, more generally, using signs to express one's mental state, requires cognitive skills to begin with. Learning a language, or learning to master any symbolic practice, changes our cognitive makeup. The symbolic system itself becomes a

causal factor in explaining the mind. If we seek to explain the difference between an English-speaker's expressing their hunger by saying "I am famished" and a Frenchman's expressing the same mental state by saying "J'ai grand faim en ce moment", we point toward their having learned different languages. If we seek to explain the difference between a human being's expressing their hunger and a dog expressing its hunger, we point toward the human being's having learned *any* language. And in many cases, we are dealing with mental states which arise from partaking in symbolic practice, such as being afraid of a stock market crash – a mental state that dogs cannot share (even though they plausibly can have some mental attitudes toward the consequences, such as starving). Here, partaking in symbolic practice such as having learned a language becomes a causal factor in explaining the mind; and if, like Searle and Grice, we invoke examples which require linguistic expression, then the mental states we will end up pointing toward in explaining the intentionality of the symbols that are used will necessarily have to include mental attitudes toward symbols in the first place. (Of course, none of this is to exclude that there can be intentional mental states, such as desiring food, which can precede the use of symbolic systems. However, these can do nothing to establish mental primacy in cases that do in fact deal with such use.)

While denying that these examples establish mental intentionality as primary, I do not intend to pursue establishing the primacy of symbolic representation either. Rather, I think the proper view is, at least when examples such as these apply, to see mental and symbolic intentionality as interlocked in the way just described: Symbolic intentionality depends on cognisers, and what cognisers do often depends on there being symbolic systems. A theoretical upshot of the two forms of intentionality being interlocked is that intended meaning and the form of intentional mental states that are expressed in the way these examples illustrate (namely linguistically) cannot be ascribed separately. (For, based on what I have said so far, one might think that we could settle the question whether someone is a competent practitioner of symbolic actions before tackling the task of ascribing mental states to her.) I already pointed out that there are in fact two mental states underlying the utterance "J'ai grand faim en ce moment": namely the hunger, but also the intended meaning ("I mean to express my hunger by this utterance"). Yet, these two separate cognitive states are expressed in but *one* observable action, namely the utterance. Borrowing from Donald Davidson's terminology: symbolic and mental intentionality are two aspects forming one vector, and the two aspects are only ever observable indirectly by observing the vector. Thus, "[i]t makes no sense to suppose we can first intuit all of a person's intentions and beliefs and then get at what he means by what he says. Rather we refine our theory of each in the light of

the other" (Davidson 1980: 258). What the assignment of meaning to symbolic actions and what intentional psychology jointly attempt is to untangle such vectors using theories whose operative structure consists in assigning intended meanings and intentional mental states. I will elaborate on its methodology and theoretical foundation in greater detail in section I.7.4.

So far, I have established that meaning something symbolically means having an intention to use a symbol in a certain way. I have also suggested that symbolic intentionality is constitutive for some form of mental intentionality insofar as learning to refer to objects in a certain way enables us to have attitudes about them in the first place (compare II.3). Yet, these two points do not by themselves establish that mental intentionality is symbolic. The latter view is introduced by Cummins as follows:

> "Haugeland (1985) credits Hobbes with being the first to have an inkling that mental representations might be language-like symbols. This is now the orthodox position, insofar as there is such a thing. (…) [I]f mental representations are symbols, then mental representation cannot be founded on similarity; symbols don't resemble the things they represent. The great advantage of symbols as representations is that they can be the inputs and outputs of computations. Putting these two things together gives us a quick account of the possibility of thought about abstractions. When you calculate, you think about numbers by manipulating symbols. The symbols don't resemble the numbers, of course (what would resemble a number?), but they are readily manipulated.
>
> Connectionists also hold that mental representations are symbols, but they deny that these symbols are data structures (i.e., objects of computation). In orthodox computational theory the objects of computation are identical with the objects of semantic interpretation, but in connectionist models (at least in those using truly distributed representation) this is not the case. Connectionists also typically deny that mental symbols are language-like. This is not surprising; given that the symbols are not the objects of computation, there would be no obvious way to exploit a language-like syntactic structure in the symbols anyway" (Cummins 1991: 6).

I can afford to be silent on the dispute between orthodox computationalists and connectionists. What my claims require is to hold that mental states, in order to be representational, need to satisfy conditions of symbolic ascribability. That is, they generally need to be states which process inputs and produce outputs (i.e. which are functionally characterised, cf. Lewis 1972: 204, 207 f.) in a way that warrants intentional ascriptions. Practically, states fulfilling such requirements can have a widely varied nature and can be implemented in an indefinite amount of ways. Theoretically, what is important is that the

states have symbolic properties. In intentional psychology, mental content is symbolic because its ascription must obey such requirements; this is the line of reasoning I am going to pursue in I.6 and I.7. In other cognitive sciences, representations are symbolic due to considerations of their functionality (this is what I am going to explore in the second chapter). While my ultimate goal is to establish connections between intentional psychology and neuroscience, I will assume no substantial connection between different forms of representation or intentionality from the get-go, since, as Cummins also succinctly notes,

> "to understand the notion of mental representation that grounds some particular theoretical framework, one must understand the explanatory role that framework assigns to mental representation. It is precisely because mental representation has different explanatory roles in "folk psychology", orthodox computationalism, connectionism, and neuroscience that it is naive to suppose that each makes use of the same notion of mental representation" (Cummins 1991: 12 f.).[19]

As far as Fodor's LOT goes, I suggest we take it with a grain of salt: The mind is indeed sometimes structured in a linguistic way, but this shouldn't come as a huge surprise given my suggestion that symbolic systems can be causes for our cognitive constitution. If we are apt at those which are linguistic, if our mind has learned to operate in and on their structure, then it seems plausible to think that at least the part of the mind that operates this way adopts a sentential structure. Yet, neither mental nor symbolical intentionality depends on its being structured so. Compare Dennett:

> "Of course sometimes there are sentences in our heads, which is hardly surprising, considering that we are language-using creatures. These sentences, though, are as much in need of interpretation via a determination of our beliefs and desires as are the public sentences we utter. Suppose the words occur to me (just "in my head"): "Now is the time for violent revolution!"—did I thereby *think* the thought with the content that now is the time for violent

[19] Cummins also urges us not to confuse mental representation with intentionality (ibid. 13-15). By that he means to warn us that representation in different cognitive sciences need not work like representation in intentional psychology or inherit the latter's kind of representations or explanatory schemes. However, throughout this book I will use "intentionality" to simply refer to all kinds of representational properties, while still sharply distinguishing between kinds in intentional psychology and representational entities in other sciences. That is, unlike Cummins, I will not (terminologically) equate intentional states with propositional attitudes (ibid.: 14), but I substantially agree with him in holding that the problem of representation in cognitive science isn't merely the problem of figuring out the nature of propositional attitudes: "anyone who assumes, for whatever reason, that a theory of mental representation must give us intentional contents (e.g., objects of belief) is making a very large assumption, an assumption that isn't motivated by an examination of the role representation plays in any current empirical theories. After all, it isn't belief of any stripe that most theoretical appeals to mental representation are designed to capture" (ibid.: 15).

> revolution? It all depends, doesn't it? On what? On what I happened to believe and desire and intend when I internally uttered those words "to myself"" (Dennett 1987: 93).

Here, Dennett insinuates two important things: firstly, the real cause for there sometimes being sentences in our head is that we are language-users (not vice versa). Certainly, this leaves open the possibility of language being naturally suited to the expression of our minds – i.e. that our minds have natural properties which lend themselves to verbal expression better than to, say, pictorial expression (for one, sounds are usually easier for our bodies to produce without additional tools than pictures). But we also know that our minds can adapt to all sorts of different forms of symblic systems – pictorial, numerical, musical, visual, and so on. So the primary explanatory cause for our sometimes having sentences in our heads such as "Now is the time for violent revolution!" is that we commonly traffic in languages – an ability for which the natural structures underlying our minds can provide necessary, but not sufficient conditions. The structure of our minds must enable us to learn partaking in symbolic practice such as language, but the learning process itself very much shapes the structure of our minds too. What happens when children learn languages? Their minds are shaped from what we might idealisingly call a "natural" state toward a decisively socially influenced state: the state of being apt at using language. It remains controversial to what degree innate structures lead us toward language-use (for contemporary nativist views see Carruthers, Laurence & Stitch 2008: esp. ch. 11 and 12), but what cannot be denied is that learning is necessary for speaking a language. And even beyond childhood, we are still constantly learning in some way or another. Many of the properties of symbolic systems are young in evolutionary terms; so it would be absurd to hold that they could all be innate. So, while it is plausible to assume that a general innate learning mechanism is usually part of our minds, it probably isn't a mechanism that specifically enables us to learn English, Quantum Physics or C++.

While this line of thought captures a lot of what I think is true about "sentences in the head", it does not fully do justice to LOT. In fact, the part of LOT that has to do with mental sentences is best understood as concerning syntax, while we are currently concerned with semantics (and it is the semantic point which is reflected in Dennett's quote). In brief, LOT is true if the objects in our minds have a syntax akin to natural language, which, at its most basic, boils down to compositionality, or the claim that "the semantic value of thought (/sentence) is inherited from the semantic values of its constituents, together with their

arrangement" (Fodor 2001: 6).[20] For example, in the case of a mental state that can be expressed by the propositional attitude "believing that all bears in Bavaria have brown fur", LOT would roughly hold that the underlying cognitive state is a composition of the cognitive entities corresponding to the concepts "bears", "fur", "brown" and "Bavaria" (plus relevant quantifiers and connectives).[21] The reason for inferring mental compositionality from linguistic compositionality is, briefly, that sentences are used to express thoughts, so if sentences are compositional, thoughts should be too (cf. Fodor and Pylyshyn 1988).

However, there are several forms of cognitive encoding which would preserve compositionality of computed representations (and thereby the productivity and systematicity of language). Smolensky, Legendre and Miyata (1992: esp. 41-45) have shown this by proving an equivalence between a parser written in TPPL, a LISP-like language that uses classical representations and a connectionist network using fully distributed representations. They conclude that "[s]ymbols and rules (…) play essential roles in (…) [explaining] crucial properties of higher cognition, but they do not play a role in algorithms which causally generate this behavior" (ibid.: abstract). Given this result, we can see that it is not necessary for there to be a one to one correspondence between constituents of thought and constituents of language in order to show how language can be used to express thoughts. By turning sentential cognitive structures into a possibility rather than a necessity, Fodor's and Pylyshyns claim becomes an empirical hypothesis. That our means of expressing our thoughts are linguistic cannot by itself settle the question whether the structures underlying our higher cognitive capacities are best described by sentential, connectionist or other means (for further bridge-building between symbolicist and connectionist representations see McCulloch & Pitts 1943 and Leitgeb 2003). Rather, it seems much more plausible that our means of expressing them is sentential because we have learned to use sentential symbolic structures, not because they are primarily sentential "in the head" – and thus, we're back at Dennett's quote.

LOT should not be mistaken for an answer to the question where intentionality comes from: "Those who defend the [LOT] hypothesis are at pains to make it clear that postulating sentences in the head is one thing, and explaining how those sentences get their meaning – giving a 'semantics for the language of thought' – is quite another" (Crane 2001: 346). But

[20] Fodor also makes a stronger claim: "In fact, (and this is no small matter) the connection that compositionality imposes on the relations between the possession conditions of concepts and the possession conditions of their hosts goes *in both directions*. That is, compositionality requires not just having the constituent concepts is sufficient for having the host concept, but also (and more obviously) that having the host concept is sufficient for having its constituents" (ibid.: 9).

[21] Personally, I find it easiest to make sense of this claim by replacing "cognitive" with "neural", but since we can find both the terms "mental" and "neural" to be in use when making this claim, I opted for the ambiguous "cognitive" here to capture the ambiguity inherent in the claim itself.

where, then, do semantics come from? One answer to this question, which is regularly combined with LOT, is that the mind is a syntactic engine driving a semantic engine; that is, that it carries out syntactic operations on structures that have semantic content (such as sentences). As Ned Block puts it,

> "the idea of the brain as a syntactic engine driving a semantic engine is (...) that we have symbolic structures in our brains, and that nature (evolution and learning) has seen to it that there are correlations between causal interactions among these structures and rational relations among the meanings of the symbolic structures. A crude example: the way we avoid swimming in shark-infested water is the brain symbol structure '*shark'* causes the brain symbol structure '*danger'*. (...)
> [Its] processors "know" only the "syntactic" forms of the symbols they process (e.g., what strings of zeroes and ones they see), and not what the symbols mean. Nonetheless, these meaning-blind primitive processors control processes that "make sense" – processes of decision, problem solving, and the like" (Block 1995b: 397 f.).

Since it is thusly assumed that "if you take care of the syntax, the semantics will take care of itself" (Haugeland 1981: 23), all the brain really has to do is carry out these computations in order to accomplish all the wondrous tasks we commonly associate with rational thought. This picture is exceedingly attractive, given the widespread and rather basic assumption that our brain's main contribution to cognition is its computational powers, and that its computational properties can be described naturalistically (i.e. non-representationally, see section II.2), thus effectively setting up a program to naturalise rational thought. However, the only way for me to reconcile LOT with my own view that semantics is not derived from internal "mental meanings", but rather crucially depends on a specific connection to the environment (see sections I.8 and II.6), is to take LOT as a hypothesis about how the cognitive bases for intentional mental states are implemented (namely as a specific form of syntactic operations). I do share its basic assumption that those parts of the brain relevant for cognition are fruitfully conceived of as engines which operate on entities that have semantic properties, but I see no way to properly construe these neural operations as sufficient for construing meaning, much less point to them in order to clear up questions about symbolic intentionality. Given this view, some of Fodor's stronger theses, such as that the semantics of English are properly researched in terms of the semantics of thought (cf. Fodor 2001), must seem outlandish.

The bottom line is that, while LOT amounts to a theory about how a physical system such as the brain can deal with the productivity and systematicity of language on the one hand and the semantic properties of symbolic systems on the other, this is far from establishing that matters of the brain can be invoked to explain semantic properties. In fact, what lends the "syntactic engine" view its explanatory force is that matters of semantics do not have to be settled by investigating the physical bases underlying mental processes in the first place. Rather, it establishes that a physical system's conforming to rational demands need only presuppose its running syntactical operations. That the semantics "will take care of itself", as Haugeland stated, should not be taken to mean that semantics are already implied by certain syntactic operations, but rather that *if* there are semantic relations, then there can be physical systems carrying out syntactical operations preserving them. As Block pointed out, the processors carrying out the syntactic operations do not need to "know" what the symbols mean. But that is not to say that semantic properties can be disregarded in favour of, or reduced to, syntactic operations, but rather the opposite, namely that what the symbols mean is not determined by the syntactic operations alone, but rather by the function that the execution of these syntactic operations implement. Evolution and learning has caused our brains to carry out specific operations (such as associating shark-representations with danger-representations), but only because they serve specific functions; and characterising the functions at least partly requires characterising what goes on beyond the brain (namely, stating the context of evolution or learning). I will pursue this view further from sections II.6 onward.

Dennett insinuates secondly that if there are sentential entities with intentional properties in our heads, then what they mean is subject to the same form of interpretation as in the case of non-mental sentences. While this is consistent with Fodor's version of LOT, highlighting the fact that "mental sentences" are slaves to the same dynamic of the underlying ascription of psychological states as non-mental sentences serves to bring out the way our ascribing psychological states and intended meanings to one another is interlocked. Crucially, the Davidsonian view that meaning is the operative part of such psychological theories, which I'm partial to (see I.7.4), does not imply that questions of psychology are explanatorily primary to our use of symbols. It is fully consistent with my claim that the most plausible way to connect symbolic intentionality and mental intentionality is to say that it takes cognitive skills to use symbols, and that there are no symbols without cognisers (here, we can embed Gricean pragmatics). Furthermore, it is consistent with holding that at least some

psychological states are intentional because they are attitudes taken toward things that inherently have symbolically referential properties (such as, but not limited to, propositions).

I.4.5. Rich Versus Sparse Representations

How should we deal with mental states which we want to call intentional, but which do not plausibly require symbolic intentionality? There are two kinds of mental states to deal with here: Mental states whose having relies on mastering concepts, and mental states which don't. For example, could anyone believe that there's a red flower on the meadow if this person were to live in a world in which no symbols of any form existed? Having such a belief requires having the appropriate concepts (*red, flower, meadow*), which in the classic philosophy of mind are often taken to require linguistic (or at least quasi-linguistic) skills (cf. Davidson 2001b: 95-105). Accepting this view implies that without symbolic practice there can be no concepts.

But concepts need not be linguistic, and if we do not want to concede that having concepts requires having a language, maybe we could employ a definition of concepts which is satisfied by multimodal/multisensory integration (cf. Holmes et al. 2009, Reig & Silberberg 2014, Deroy 2015): Perhaps the most important feature of concepts is that they can be applied across different relevant instances – such as the concept "flower" applying to different flowers – and so we should hold that our concept "flower" could also consist in a mental entity that bundles all the different modalities needed to react consistently to flowers, including any such bundling that would be formed based on past interaction with flowers (i.e. knowledge or memory of flowers). Such multimodal concepts go a long way to explaining our behaviour toward flowers; perhaps we have learned from interaction, perception and/or imitation that they're good for picking.

But in this story, what role does intentionality play? It seems sufficient to tell it in terms of behaviour, of causes and interaction, as I have just done. Adding that someone who has, in this sense, learned to properly interact with flowers has a mental state that is directed at flowers does not seem to have any explanatory surplus compared to the non-intentional explanation. Sure, we might *like* to say that some of his mental states are directed to flowers. Then again, we might mostly want to do so because it is natural for us to take the intentional stance, i.e. for us to ascribe intentional attitudes in order to explain behaviour (cf. Dennett 1987); and our taking this stance does not require that the cogniser whose state we are

describing intentionally is, in fact, an intentional agent – much like describing a thermostat as believing that the room is too cold does not imply that the thermostat satisfies any substantial criteria for being an intentional agent (see section I.7.5). In these cases, it is sufficient to use only a sparse notion of "representation", namely a purely causal one. We should distinguish it from a richer notion which implies that cognisers are generally moved by the laws of intentional psychology, or, in other words, minds whose features are causally explained by intentional features: Features such as learning to partake in symbolic practice, learning to read, learning to conform to social standards, learning to meet institutionalised expectations, learning that giving a promise means intending to keep it, and so on.

> "What's special about us is that we don't just do things for reasons. Trees do things for reasons. But we represent the reasons and we reflect on them, and the idea of reflecting on reasons and representing reasons and justifying our reasons to each other informs us and governs the intentional stance. We grow up learning to trade reasons with our friends and family. We're then able to direct that perspective at evolutionary history, at artifacts, at trees. And then we see the reasons that aren't represented, but are active" (Daniel Dennett in Edmonds & Warburton 2015: 130).

Explaining behaviour related to said kinds of higher cognitive skills is what the true explanatory value of intentional psychology consists in, and if explanations can do without it, then there is no good reason to take the use of semantic terms such as "is about" or "refers to" as implying that we are dealing with the explanation of a genuinely agential/intentional phenomenon – even if we are still free to take the intentional stance toward it, as in the case of the thermostat.

For example, if I buy a pack of mozzarella, and am motivated to put it into the fridge in an upright position, then the cause (i.e. the proper referent of an explanation of what goes on in my mind) are representational properties of the pack of mozzarella. The only way to determine whether the pack stands upright is to know the difference between the typeset "Mozzarella" and "Mozzarella". Of course there are physical, biological and anatomical prerequisites to meet in order for me to acquire this knowledge: I have to be a kind of organism which can be conditioned to distinguish "Mozzarella" from "Mozzarella", which can generalise over a potentially infinite class of type sets in order to determine whether they're upright or upside down, and which can be motivated to act in accordance with this knowledge. Still, there is nothing which all upright typesets have in common which could be explained in purely physical terms (compare Fodor's pointing out that "interesting

generalisations (...) can often be made about events whose physical descriptions have nothing in common" in his 1974: 103), or determining which could be traced back to biological and anatomical configurations or disposition. Thus, in cases such as this, the explanatory cause always and exclusively consists in a representational feature.

To be sure, both sparse and rich notions of representation can be described as phenomena of mental directedness. Appetite for chocolate, fear of snakes, fear of stock market crashes, appreciation of mathematical proofs, and so on – all of these mental states have intentional objects, even though the former pair can be sparse and the latter can only be rich. What makes the difference is that it is only the rich notions which come with the notorious problems which intentionality poses for the cognitive sciences (and for naturalistically inclined philosophers of mind), namely phenomena of normativity and rationality. Only in cases of rich intentionality are norms and matters of rationality necessary parts of causal explanations of mental states. Desiring to come up with an elegant mathematical proof, composing classical orchestra pieces, seeking to be a virtuous person, protesting nuclear power, and so on – if we wish to explain the respective underlying cognitive states, standards of rationality which are external to the cognisers will have to be evoked as causes of mental effects: Explaining why someone speaks English is not merely a matter of explaining their internal cognitive state, but also of specifying English grammar as an external cause of this cognitive state. In these cases, representations (such as a rulebook of English grammar) are invoked as causal determinants of behaviour (compare section II.3). Typically, it is in the interaction between cognisers and representations where we can locate matters of rationality and normativity: namely, by the cogniser's adhering to what is specified representationally, by adhering to logical, social, moral norms and rules. Clearly, such cases differ from being hungry or fearing snakes, which we can explain without invoking representations as causes.

To be sure, I do not wish to claim that rich or high-level cognition, such as skills required for making informed political election choices, are typical for human cognition across the board (and we may even imagine being able to live out most of our days without invoking them). But I do claim that it is characteristic for human cognition to *include* such high-level cognitive skills and that those who cannot have them appear stunted. So, an account of the mind which disregards rich representation cannot amount to an account of the human mind.

It should also be noted that the difference between sparse and rich notions of representation does not coincide with the difference between being free of and being causally

influenced by external restrictions. Rather, the sparse notions we have been discussing are also governed by external restrictions, namely those of evolutionary adaptation: we would not be hungry and we would not fear snakes if either didn't serve an evolutionary function. Whereas rich notions are governed by aims such as being rational, being consistent, being virtuous. Both notions are *teleological* in nature, that is, they are formulated in terms of aims (from the Greek word "telos", τέλος). In sparse notions, we usually find biological or organismic functions which we can describe as aims, whereas in rich notions, we will refer to norms which are socially constituted and/or implemented (see II.7). The latter will *build* on mental capacities which are evolutionarily acquired – we can only ever build on what nature has equipped us with –, but, unlike in the case of those connected to sparse notions, they will not be evolutionarily *determined*: evolutionary and organismic explanations will not be sufficient to have and explain them.

Essentially, intentional psychology as explored in this chapter is concerned with rich notions of representations, whereas I will explore the role of sparse notions as invoked in neurobiology in chapter II. This difference in representational notions points toward a dilemma that the cognitive neurosciences face: Among other things, they seek to explain intentional states. But it seems that either the notion of representation is too sparse to support the notion of intentionality, whereby we would lose all of the explanatory power intentional psychology has, or it is not sparse enough for being meaningfully tied to neurobiology (and therefore, to potential descriptions of what fundamentally produces mental states). The most fundamental field in the neurosciences, which cognitive neuroscience has to be integrated with (which can be read strongly as "reduced to" or weakly as "communicating productively with"), is neurobiology, and neurobiology decidedly rests on physicalism. Thus, as Dretske poignantly put it, the challenge is to "bake a mental cake using only physical yeast and flour" (Dretske 1981: xi). These are some of the challenges that will be addressed in chapter II. The idea, very briefly, is that physical states which allow for cognitive processing are endowed with intentional properties. Namely, since we are cognitively outfitted to consistently associate certain environmental cues with arbitrary signifiers and to adapt our behaviour accordingly, we can use environmental cues, which are evidence for ascribing intentional mental states, to explain and predict the kind of behaviour typically or normatively connected to the mental states identified by the respective attitude and content.

I.5. Mental Constructionism

In this chapter I am going to explore the ontological consequences of my conception of mental states. Since I construe the reality of mental states as tied to their explanatory value, treating them as kinds in psychological theories, I adopt a kind of pragmatic scientific realism, which holds that what is real is determined (or at least strongly informed) by what kinds we use in scientific theories. I am not going to argue that this is the only adequate explication of what it is to be real, but that this is the conception relevant for mental states. This section will then lead into the subsequent sections discussing how mental states are used to explain phenomena. In this sense, this is a "theoretical" section discussing the ontological underpinnings of what is explored in the following "practical" section.

The view I am going to endorse in this book is a weaker form of what has been dubbed "mental constructionism". This weaker form sheds some of the stronger version's theses about representational properties being married to linguistic competence, and about the spuriousness of intentional laws – but more on this later on. Some of its (stronger) forms have been attributed to Wilfrid Sellars and Donald Davidson, among others. Both of them make a case for mental states being theoretical objects, and it is this notion of our constructing mental objects according to their explanatory value and theoretical aesthetics which gives constructionism its name. For example, Sellars argues that mental states such as thoughts and perceptions should be understood as theoretical entities postulated to explain overt behaviour (cf. Sellars 1997: 90-117). Since they can only play this role if they are intersubjectively accessible, he points out that "concepts pertaining to such inner episodes as thoughts are primarily and essentially *intersubjective*, as intersubjective as the concept of a positron, and (...) the reporting role of these concepts — the fact that each of us has a privileged access to his thoughts — constitutes a dimension of the use of these concepts which is *built on* and *presupposes* this intersubjective status" (ibid.: 107, §59; see also his 1957). And Davidson treats mental states as akin to abstract objects such as units of measurements: "Just as we cannot intelligibly assign a length to any object unless a comprehensive theory holds of objects of this sort, we cannot intelligibly attribute any propositional attitude to an agent expect within the framework of a viable theory of his beliefs, desires, intentions, and decisions" (Davidson 1980: 221, also see Field 1975). Such comparisons highlight the abstract, theory-dependent nature of mental states, as well as (in Davidson's case) a holistic view of the mind: the view that mental states pick out relative properties in an interdependent web of such states, just as metres specify points on a relative scale of size or length. For

example, for any belief to qualify as being about a tree, the belief's bearer must have "many general beliefs about trees: that they are growing things, that they have leaves or needles, that they burn. There is no fixed list of things someone with the concept of a tree must believe, but without many general beliefs, there would be no reason to identify a belief as a belief about a tree, much less an oak tree" (Davidson 2001b: 98, also compare Searle 2000: 107).

As previously noted, these relative properties are expressed in propositional form, assigning semantic content by means of a that-clause which is preceded by the specification of what is called the intentional mode (see e.g. Searle 2000: 99 and I.1). For example, you could justifiably assign to me the mental state that I believe that as I am writing this, it is way past my bedtime. In this case, "believing" is the intentional mode (others are doubting, dreaming, desiring, and so forth), and "as I am writing this, it is way past my bedtime" is the propositional content. This form of assigning mental states is not necessarily linguistic: For example, Sellars required for a basic representational state "to have propositional form, (…) [it] must represent an object and represent it as of a certain character" (Sellars 1981: 336) , a requirement met by a range of non-linguistic forms of representations, such as maps.

Just like the notion of propositional attitudes, the roots of constructionist views can be traced back to Bertrand Russell, who expressed its core back in 1928:

> "modern science gives no indication whatever of the existence of the soul or mind as an entity; indeed the reasons for disbelieving in it are very much of the same kind as the reasons for disbelieving in matter. Mind and matter were something like the lion and the unicorn fighting for the crown; the end of the battle is not the victory of one or the other, but the discovery that both are only heraldic inventions. The world consists of events, not of things that endure for a long time and have changing properties. Events can be collected into groups by their causal relations. If the causal relations are of one sort, the resulting group of events may be called a physical object, and if the causal relations are of another sort, the resulting group may be called a mind. Any event that occurs inside a man's head will belong to groups of both kinds; considered as belonging to a group of one kind, it is a constituent of his brain, and considered as belonging to a group of the other kind, it is a constituent of his mind. Thus both mind and matter are merely convenient ways of organising events." (Russell 1973[1935]: 142 f.)

Not minding the details of Russell's view, what we should take away from this quote is that the concept of "mind" is but one way to organise certain features of the world (in Russell's case, events characterised by (a) specific type(s) of causal relations). I take this view to be a natural companion to the modern scientific worldview, namely the view that, roughly

speaking, there are objective features of the world which we have more or less good access to by way of perception, measurement and various forms of interaction. Consequently, we construct scientific theories in an attempt to systematically relate these features to one another (such as by way of causal relations, identity relations, structural relations such as mereology or the forming of certain spatial or temporal patterns, and so on) in order to explain them and the phenomena by which they reveal themselves to us. In such theories, certain key concepts will serve to give structure and support to the theories and the laws they entail. In the natural sciences, natural kinds serve as such key concepts. Conceptually, they enable us to distinguish between generalisable or universalisable relations and non-universalisable ones.[22] Similarly to what Russell insinuates, psychological explanation depends on taking causal relations to be of a specific sort: psychological explanation is a specific form of causal explanation, one which relies on psychological laws and psychological kinds. I will go into more detail regarding this kind of explanation in section I.6.

The constructionist view as has just been laid out has been dubbed "antirealist" and contrasted with a "realist" view of the mind, where mental states are taken to be as real as "tables, stones and electrons" (Heil 2000: 131). However, the mere fact that intentional psychological explanation traffics in abstract terms and depends on a theory is not to say that it assumes an antirealist outlook. On the contrary: a compelling view advocated by naturalistically inclined philosophers is that scientific theory itself is the measure of what is real: "in the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not" (Sellars 1997: 83, §41; compare also Quine 1980: ch. 1). So, what we should be looking for to judge realness is not a pretheoretic notion of the concept "mental state", but sound criteria for determining whether these concepts capture real causal relations, and it is in virtue of capturing these that they figure in psychological theories in the first place. (I will go on to argue in section I.6 that this criterion is in fact met by intentional psychology.)

And following Dennett, we should construe realism of theoretical objects as their being *good* theoretical objects. For instance, rather than pursuing a purely metaphysical question about whether theoretical/abstract objects such as centers of gravity are real, "we should be (…) more interested in the scientific path to realism: centers of gravity are real because they are (somehow) *good* abstract objects. They deserve to be taken seriously,

[22] I say "conceptually" because it is not all that clear whether there can be a purely empirical or epistemic criterion for distinguishing between laws and non-lawlike regularities, and many believe that the notions of laws, natural kinds, induction etc. are essentially circular (see e.g. Davidson 1980: 217 f., Fodor 1974: 102) – even if they may form a virtuous rather than a vicious circle. See also section I.6.2.

learned about, used. If we go so far as to distinguish them as *real* (contrasting them, perhaps, with those abstract objects which are *bogus*), that is because we think they serve in perspicuous representations of real forces, "natural" properties, and the like" (Dennett 1991b: 28 f.). Or, as Richard Healey worded Dennett's "deep pragmatist point – it is more important to appreciate the purposes for which we agents created concepts of these things than to undertake the quixotic and ultimately unrevealing task of relating them in an orderly way to some allegedly fundamental ontology" (Healey 2013).

Sticking with Heil's paradigmatic instantiations of "realness" such as tables, stones or electrons, we should hold that mental properties are perhaps closest to electrons, insofar as these are construed as theoretical entities as well; that is, they play a central role for the explanatory value of these theories. Whereas tables and stones can be defined quite independently of being kinds in a theory, insofar as tables can be (at least tentatively or heuristically) defined by their function, and stones can be defined by being made of a certain material. Ultimately, tables and stones will stand in some relation to theories (tables more to social and cultural theories, and stones more to physical, geographical or architectural theories), but their everyday use diverges from their being used as kinds in theories. That is, even if "being made of stone" works as a kind in architectural theories, there is no requirement for any of us non-architects to (even implicitly) refer to architectural theories when talking about stones. Whereas it seems quite impossible to talk about electrons without presupposing that they play a major role in the explanatory value of physical theories, just as it seems impossible to talk about semantic and mental properties disregarding their analogous theoretical foundation. Davidson and Fodor support this train of thought: "The analogy with physics is obvious: we explain macroscopic phenomena by postulating an unobserved fine structure. But the theory is tested at the macroscopic level. Sometimes, to be sure, we are lucky enough to find additional, or more direct, evidence for the originally postulated structure; but this is not essential to the enterprise. I suggest that words, meanings of words, reference, and satisfaction are posits we need to implement a theory of truth. They serve this purpose without needing independent confirmation or empirical basis" (Davidson 2001a: 222). Furthermore, "the [commonsense psychological] theory's underlying generalisations are defined over unobservables, and they lead to its predictions by iterating and interacting rather than by being directly instantiated" (Fodor 1989: 7).

Again, the mere fact that these objects are theory-dependent and not directly observable should not sway us to doubt their ontic reputation, for even

> "a properly trained physicist, who can respond systematically differently to differently shaped tracks in a cloud chamber will, if responding by non-inferentially reporting the presence of mu mesons, count as genuinely *observing* those subatomic particles. The physicist may start out by reporting the presence of hooked vapor trails and *inferring* the presence of mu mesons, but if the physicist then learns to eliminate the intermediate response and respond directly to the trails by reporting mesons, the physicist will be observing them" (Brandom 2002b: 96).

Like Sellars, Brandom holds that theoretical and directly observable objects do not differ in kind, but only in whether we access them inferentially or non-inferentially (cf. Brandom 2002a: 362). Consequently, merely pointing out that being able to directly or non-inferentially observe something – a table, a stone – will not ontologically separate it from theoretical objects whose observation we take to be inferred, namely by invoking a theory which specifies under which conditions an objects counts as being observed, present, instantiated or existing.

So, while there may be no ontological divide between my form of constructionism and straightforward mental realism, a notable difference pertains to whether there is the possibility of finding pretheoretical or purely subjective, internal mental states. What I wish to rule out by adopting constructionism is the notion that an immediate acquaintance with mental states is either necessary or sufficient for our knowledge about them (see section I.7). For some form of immediate acquaintance seems to me to be the only way to escape the view that mental terms are, to some degree, abstract and theory-dependent. However, allowing immediate acquaintance to play a fundamental role does not seem too appealing to begin with, for not only does it bring with it the bane of solipsism (see I.7.2), it also rules out all forms of externalism, some of which will later turn out to be intrinsic to my view. Of course, none of this implies a denial of "qualia", qualitative experiences which are essentially internal and subjective (see section I.9.4). I simply deny that subjective experience is all there is to be said about the realm of mental states.

Given the fact that theories in contemporary cognitive science are decidely committed to an at least broadly construed physicalist ontology (i.e. assuming at least a general consistency with physics, and often also the hope of ultimately reconciling non-physicalist ontologies with physicalist ontologies), mental constructionism is an adequate position to hold. Constructionist intentional psychology is consistent with a basic physicalism, that is, the notion that all real things are describable by physics: at their most basic, all objects should be in some form consituted by elementary particles, forces, natural properties of space and time and the like, which form an inventory of natural kinds in physics. The idea that objects which

are constituted by physical kinds, and some properties of these objects, can also figure as kinds in laws beyond physics (such as psychological laws), without assuming that these laws relate to physics in an interesting way (such as being restatable or reducible to physics, eliminable by physical laws, etc.) is compatible with accepting basic physicalism. Reducibility and eliminability are options, but not implied.

While we commonly also describe many macroscopic objects as "physical objects", such as tables, paintings, or human beings, we should not mistake this for claiming that these are themselves objects in physicalist ontologies. Rather, we should take it as shorthand for referring to their material properties (as opposed to immaterial objects, such as ideas, mathematical proofs, or ideal beauty). While tables, paintings and human beings may be material things, they are, according to physics, not among the realest things our universe has to offer: they are not what figures in the most basic of physical laws. In other words: They aren't natural kinds in physics. The distinction between material and immaterial only presupposes physics insofar as material objects are made up of physical particles by way of composition – a composition that is defined by physical notions such as forces, space and time. (Although, taking the stance of a quantum physicist, we would have a hard time finding the boundary between a table and its surroundings on the quantum level – still, any point in space within the boundary should turn out to be such on the microscopic level that it allows for the table having its table-esque qualities on the macroscopic level). There are, of course, various theories about how immaterial things are or could be composed of physical objects, but these are prone to stir controversy (and here we might want to distinguish between immaterial things such as mathematical proofs, whose existence at the very least does not violate physical laws, and such things as ghosts or angels, which on many common accounts would). For example, a rough idea for how ideas could be composed of physical objects could hold that ideas are things that are both mental and social, and that mental things are composed of persons and the relations between their neural mechanisms and their environment.

Certainly, physicalism offers straightforward mental realists an easy way out by urging to just identify mental states with brain states and be done with it – even in the absence of empirical proof about the identity of any specific mental state with a specific brain state. Don't we readily identify many macroscopic objects with their physical microstructure, even in the absence of knowledge about how one exactly relates to another? Would adopting the view that thoughts are composed of elementary physical entities amount to a greater leap of faith than the view that chairs are? If, in a physicalist framework, we so readily assume the latter, why not the former? There are two important reasons: Firstly, chairs have all sorts of

properties which promise to readily be reducable to more basic physical properties of elementary chair-parts; and some chair-properties can be neatly integrated into physical laws. Whereas mental states have some odd properties which can't, at least not in an obvious way – one of them being that thoughts have intentional content. This is not to say that intentional content must seem like an odd phenomenon – it certainly doesn't when invoked in the framework of intentional psychology –, but intentionality can *seem* odd when viewed from a physicalist point of view (see II.2).

Secondly, while we generally acknowledge that objects such as chairs are identical with their physical components, we cannot say the same about all lawlike generalisations in which such objects figure. In the case of chairs, explanatorily valuable statements such as "chairs are made by carpenters" or "chairs are for sitting" make much less sense when referring to the physical description of a chair instead of chairs per se. Fodor made this point exceedingly clear for the case of money:

> "Suppose, for example, that Gresham's 'law' [in economics] really is true. (If one doesn't like Gresham's law, then any true generalisation of any conceivable future economics will probably do as well.) Gresham's law says something about what will happen in monetary exchanges under certain conditions. I am willing to believe that physics is general in the sense that it implies that any event which consists of a monetary exchange (hence any event which falls under Gresham's law) has a true description in the vocabulary of physics and in virtue of which it falls under the laws of physics. But banal considerations suggest that a description which covers all such events must be wildly disjunctive. Some monetary exchanges involve strings of wampum. Some involve dollar bills. And some involve signing one's name to a check. What are the chances that a disjunction of physical predicates which covers all these events (i.e., a disjunctive predicate which can form the right hand side of a bridge law of the form 'x is a monetary exchanged...') expresses a physical natural kind? In particular, what are the chances that such a predicate forms the antecedent or consequent of some proper law of physics? The point is that monetary exchanges have interesting things in common; Gresham's law, if true, says what one of these interesting things is. But what is interesting about monetary exchanges is surely not their commonalities under physical description. A natural kind like a monetary exchange could turn out to be co-extensive with a physical natural kind; but if it did, that would be an accident on a cosmic scale" (Fodor 1974: 103 f.).[23]

Mental constructionism honors the defining role kinds in the special sciences play in lawlike explanation without contradicting physicalism, insofar as it holds that mental states

[23] But see Boyd 1999 and Jones 2004 for criticisms of this view.

are theoretical constructs invoked to explain certain physical events whose occurrence can be explained by invoking intentional kinds. These are not themselves explanatory kinds in physical ontologies, but if to exist is to be a bound variable in one of our best scientific theories, as Quine would have it (cf. Quine 1980: ch. 1), then *of course* mental states are real, since they are kinds in one of our best theories to explain some physical events, namely human behaviour. What I urge is to expand the notion of what our "best theories" are by covering those which have great/indispensable/pragmatic explanatory value (for the details, see section I.6.2). Then, scientific realism implies that mental states are very much real, in the sense that to give an explanation in terms of mental states is to employ a theory that captures real causal relations.

This picture is in line with how we usually take many things to be real objects, quite independently of what physics say about them. For instance, determining how an object is composed of physical microparticles is completely beside the point when it comes to settling the question whether something is in fact a table, a painting, or a human being. For something to be a table, it is usually enough for it to usually function as a table – for sitting around it during social gatherings, for putting food on it, and so on. Once color has been applied to a canvas using a paintbrush, once that canvas hangs in a museum, and once we judge that it has an aesthetic quality to it, then it has long been decided that it is a painting. And usually, someone's behaving in a human fashion says all we need to know about deciding whether we are dealing with a human being. There are exceptions in each case, but particle physics rarely ever have anything to do with it. (We could dream up such cases, but they would have no bearing on the issue at hand.) Usually it doesn't matter in the least whether we are dealing with the most basic building blocks of our reality when dealing with tables, paintings and human beings. That is, in all cases in which we can question the reality of these objects, and whenever we stress that a table, a painting or a human being is a "real object", we point to the fact that the object in question has not been made up, imagined, hallucinated, or so on. But it has little to do with committing to the fact that these objects are the building blocks of our physicalist ontology. Rather, it has to do with the fact that invoking them is the right way to provide the explanations we seek:

> "The problem is (...) think[ing] that if you give the lowest-level atomic explanation, then you have given a complete account of the causation: that's all the causation there is. In fact, that isn't even causation in an interesting sense. (…) The problem with that is that it ignores all of the higher-level forms of causation which are just as real and just as important. Suppose you

> had a complete atom-by-atom history of every giraffe that ever lived, and every giraffe ancestor that ever lived. You wouldn't have an answer to the question of why they have long necks. There is indeed a causal explanation, but it's lost in those details. You have to go to a different level in order to explain why the giraffe developed its long neck. (…) If I want to know why you pulled the trigger, I won't learn that by having an atom-by-atom account of what went on in your brain. I'd have to go to a higher level: I'd have to go to the intentional stance in psychology. Here's a very simple analogy: you've got a hand calculator and you put in a number, and it gives the answer 3.333333E. Why did it do that? Well, if you tap in ten divided by three, and the answer is an infinite continuing decimal, the calculator gives an 'E'. Now, if you want to understand which cases this will happen to, don't examine each and every individual transistor: use arithmetic. Arithmetic tells you which set of cases will give you an 'E'. Don't think that you can answer that question by electronics. That's the wrong level. The same is true with playing computer chess. Why did the computer move its bishop? Because otherwise its queen would have been captured. That's the level at which you answer that question" (Daniel Dennett in Edmonds & Warburton 2015: 126 ff.).

What we should do, then, is point out that these objects, insofar as they are real objects, are important for specialised ontologies; i.e. ontologies of special sciences (compare Fodor 1974). Tables are paradigmatic objects in certain crafts, paintings have a special bearing on crafts and sciences surrounding art, and human beings figure as important objects in anthropology, biology, and, to some degree of abstraction, in many other sciences which deal with characteristically human features (such as psychology, sociology, economics, and so on). Thus, if we want to take serious the notion that it is not just physics which offers us real things, we should concede that to be an object, and to be real, is to be a theoretical term in any one good theory in any given scientific domain. Therefore, I adopt the view that in order for things to be real, they should figure in good theories; but good theories exist well beyond physics. If mental states turn out to be essential to good psychological theories, and psychological theories in general help us explain, understand and/or predict what's going on in our world, then mental states are real objects. Physicalism is not contradicted insofar as accepting non-physical kinds does not rule out that any object instantiating such a kind is also describable physically. Being describable physically and being an explanatory kind in physics is obviously not the same. So, some explanatory value a non-physical kind has could be lost under a physical description. For example, chemical kinds retain all their explanatory power when stated in physical terms, but economic ones do not. And perhaps it is true that ballet *is* particle physics; but for all we know, it isn't prudent to expect explanations of dance moves in

terms of particle interactions. So, it sometimes simply isn't prudent to want to reconstruct certain kinds in the special sciences in terms of physical kinds, even if they can or could be. Saying that a chair isn't just an object fashioned to sit on, but that it is also describable as a physical object, reflects our optimism that ideally we could list all physical components of the chair and their interrelations, with these components and relations constituting kinds in physics, even though all things which have been fashioned to sit on need not (and plausibly are not) specifically and exclusively characterised by a physical kind term, but rather by a disjunction of such terms which have no explanatory property in common that could be characterised in physical terms (compare Fodor's previous quote about Gresham's law). Under a physical description, any explanatory value the chair has in, say, woodcarving, will be lost.[24]

One position which honors these insights and allows for a plurality of descriptions of the same event (say, as both mental and physical) is Davidson's view (cf. Davidson 1980: 207-227). I will adapt it in the following way: Physicalism does not require each object to itself be reducible to a physical object, but it rather requires that each event whose description invokes non-physical objects (i.e. objects which are denoted by a kind term in a theory other than physics) be also describable as a physical event (speaking in Davidsonian terms for the case of the mind-body relation: that each mental event be token-identical to a physical event). For example, an exchange of things which have similar monetary value is under at least one description a physical event, but having a certain monetary value is not a physical property.

I have stressed that what settles ontological matters surrounding mental objects is the role they play in robust theories. And yet, there is a crucial difference between mental states and other theoretical terms such as centers of gravity (Dennett 1991b: 27 ff.): While the latter can be taken to describe, explain or predict phenomena which are independent of the theories about them, there is one sense in which mental states do fundamentally depend on theories about them. To make this point, let me first say a bit about what a theory is. A theory about mental states is a systematised set of intentional psychological laws such as "if A expresses her sincere belief that she will attend the conference tomorrow, then, with a certain probabilistic leeway depending on her reliability, she will be at the conference tomorrow", or at least a set of axioms and empirical knowledge from which laws like these follow. A theory can be more, but it cannot be less: An essential part of it is that it either consists of or implies

[24] Strikingly enough, theories themselves aren't physical objects (and possibly only in a very roundabout way can they be said to be "composed" of physical objects), but physicalism cannot do without the notion of theories: no physicalism without physics, and no physics without the notion of theories (cf. Chakravartty 2013). So even physicalism itself needs to allow for some non-physical objects.

laws such as these. And if someone has laws like these at their disposal, then they have a psychological theory. If they didn't have a theory, any generalisations like the one I just mentioned would not be a law (even if it were true), for it is the fact that they are part of a nomological theory which is necessary for their being laws and not mere generalisations (see section I.6.2).

Now, the difference between mental states and other theoretical entities such as centers of gravity is that the latter, and those effects which lead us to positing centers of gravity at certain points, do not care about our theories about them – at all. We might be in dispute about whether centers of gravity exist without our positing them (since they might merely be an abstraction that only exists as a consequence of the theory's positing them), but we do not assume that, say, two specific planetoids have a specific gravitational pull on each other only because mankind has come up with physics. Again, we can debate whether things such as a planetoid and a gravitational pull exist outside of our theories (the words certainly wouldn't exist without our linguistic practice, the concepts wouldn't exist without physics, they wouldn't mean what they do without the theories in which they are used, and so on), but whatever real event that is meant to be expressed by the statement "planetoid X has a gravitational pull on planetoid Y" is quite independent of linguistic and scientific practice. What this ultimately means is that this theory, in which planetoids figure as natural kinds, is one in which human linguistic practice does not. Some radical antirealists may still want to deny this; but that is of no consequence for the task at hand, for what I would like them to acknowledge is not that there is a world that is independent of our descriptions of it, but merely that whatever goes on in the case of centers of gravity differs crucially from what goes on in the case of mental states. The point is not whether planetoids exists independently of human theories (i.e. pretheoretically), but rather that human theories are not a natural kind in physics. Of course, such antirealists should really want to say something beyond what goes on once we're committed to the domain of physics; but that is not what's at stake here. What's at stake is that in the case of mental states, theories about them constitute a causal factor for having them – at least in some crucial cases. What matters is that the existence of at least some mental states cannot be made sense of without a theory about them.

Again, consider my example of a probabilistic psychological law: "if A expresses her sincere belief that she will attend the conference tomorrow, then she is likely to be at the conference tomorrow". In most cases in which A expresses her belief – cases which go beyond those in which she is merely asked to state her belief by, say, a psychoanalyst –, she does so to signal that she is likely to attend the conference the next day. But signalling this

can only succeed if the person she tells this has a theory which comprises of or at least implies said law. The latter person literally has no other way of grasping the significance of A's utterance than by being sensitive to certain semantic properties of it, and than employing her very own common-sense psychological theory in order to make sense of these properties.[25] In order to explain what is going on between a speaker and the interpreter of her utterances, referring to non-intentional properties of the interpreter which could alert her to what the speaker is trying to get across will not suffice. Sure, there are many interesting questions to be asked about the interpreter's physiology, how it enables her to be able to employ a common-sense psychological theory, or how it enables her to be sensitive to semantic properties in the first place. But all of these explanations can only ever begin or make sense if the interaction has been recognised *qua its semantic nature*. The speaker's behaviour, expressing her belief, would make no sense at all if the interpreter could in principle have no psychosemantic theory at her disposal. It could be that in fact the interpreter lacks such a theory, thus rendering the speaker's utterance as pointless as making a promise to a rock; but the important point is that the speaker is *justified* in expecting her to be able to employ one. And that is what we, as potential interpreters of our peers' behaviour, usually are. At the most basic, the psychological ability to have a theory of mind, and to have a psychological theory which enables us to infer someone's intention from their utterances, are preconditions for having some mental states in the first place.

I did restrict this claim to *some* mental states; and these would at least be those which are essentially interpersonally functional, such as commitments to attending conferences. That is, it only works in those cases where having a mental state is tied to the ability of expressing it and assigning it to others (see Davidson's view in section I.7.4 and I.7.5). Other mental states might make perfect sense without anyone's having learned any kind of theory, such as being afraid of snakes (see I.4.5). A mental state of this sort would still be perfectly functional if there was no one but the bearer of said fear on this planet (plus at least one snake). Allowing ascriptions like these might make me seem generous in comparison to Davidson, who argued that in fact *all* mental states presuppose an intersubjective practice; a claim which makes sense when taking into account that mental states have content, and that, given Davidson's notion of content, there cannot *be* any content without interpersonal linguistic practice (cf. Davidson 2001b: 213, for an evolutionary perspective compare Carruthers & Smith 1996: ch. 20). The prima facie weaker position here would be to assume, with Dennett (see his 1987), that the ascription of mental states, such as the fear of snakes, is perfectly

[25] Note that assuming that someone has a theory that enables them to understand semantic properties isn't the same as subscribing to the notion that theory-theory is true (cf. Carruthers & Smith 1996).

justified if the object, which said state is ascribed to, shows all signs of having this mental state. Of course, we can only take Dennett's "intentional stance" toward any bearer of mental content because we are already part of a content-assigning practice, which itself depends, as Davidson held, on our having an interpersonal linguistic practice in the first place. Thus, Dennett's position may in fact only be weaker at a first glance. I will just mark this point here and explore it further in section I.7.5. What matters for my claim for now is that at least some mental states require having a psychological theory in the sense described above. Maybe *all* of them *also* require linguistic practice – and, as Davidson held, since having a language requires having a psychological theory (see I.7.4), perhaps all mental states thus require having a theory akin to the one I labelled a psychological one. In any case, subscribing to the latter view will not be necessary for what follows.

I.6. Explanation in Intentional Psychology

I.6.1. Explanation and Ontology

The central concept in evaluating psychological theories is *explanation*. This is because, as I have argued, mental states should be understood as theoretical concepts introduced in psychological theories, and because this is so, they are weeded out according to their explanatory value. What matters, once again following Dennett's "scientific path to realism" (Dennett 1991b: 28), is that they make for "*good* abstract objects" (ibid.: 29, see also section I.5). Ontologically, this means that psychological states exist as such because there are good theories in which they figure, and these theories are good because they have explanatory value.

The concept of explanatory value itself is intensional as well as relative. It is intensional insofar as what counts as explaining something depends on the description of what is to be explained. Generally, that something is intensional means that matters of designation (or "extension") are insufficient in order to properly deal with it. Following Frege (1892), the fact that one thing, such as planet Venus, can be referred to in several ways, such as by both "Morning Star" and "Evening Star", reflect its various "senses" or intensions. Intentional (psychological) contexts are paradigmatically intensional: Since "Reginald Kenneth Dwight" and "Elton John" designate the same person (thus sharing their extension), "Reginald Kenneth Dwight composed *Crocodile Rock*" follows logically from "Elton John composed *Crocodile*

Rock". However, it is safe to assume that a lot more people know that Elton John composed *Crocodile Rock* than that Reginald Kenneth Dwight composed *Crocodile Rock*. So, in order to find out whether an individual knows whether Dwight composed *Crocodile Rock*, it is not enough to find out whether they know that Elton John composed *Crocodile Rock*, but also whether they know that Dwight goes by the name of Elton John. Here, the form of description (i.e. the intension) matters more than just the factual identity of what the two descriptions refer to (i.e. the extension).

Analogously, one description may count as an explanation, while another does not, despite their having the very same extension. To borrow an example from Davidson (1980: 17), a hurricane may explain the occurrence of a catastrophe. If this hurricane was reported on page 5 of Tuesday's edition of the *New York Times*, then "the event reported on page 5 of Tuesday's *NY Times*" refers to the same thing as "a hurricane". Still, were someone to ask: "What caused this catastrophe?", the answer "a hurricane" will surely count as an explanation, whereas "the event reported on page 5 of Tuesday's *NY Times*" would most likely not. This is because explaining something requires making it intelligible to someone, and what counts as intelligible to someone depends on what this person knows. For example, were someone to actually know that this hurricane in fact *was* (exclusively) reported on page 5 of Tuesday's *NY Times*, then, too, the answer "the event reported on page 5 of Tuesday's *NY Times*" would count as an explanation. However, we can easily imagine cases in which the lack of said knowledge would result in the unintelligibility of this answer.

This dependence on individual knowledge establishes the intensionality of the concept of explanation, accounting for why the fact whether something counts as an explanation of something else hinges on the way it is described. This is especially true in all cases of semantic explanation, that is, in elaborating on what a certain concept or term means. For example, we might be unaware of what "intensionality" means, and an explanation of its meaning would necessarily consist in giving another "sense" or description (intension) of the same (i.e. extensionally identical) concept. You can see this form of explanation at work throughout this very paragraph in my attempts to make the meaning of "intension" clear.

Of course, there are other important forms of explanation beside semantic explanation. In the example just cited, we looked at one form of causal explanation, where a hurricane was invoked to causally explain the occurrence of a catastrophe. Usually, scientific explanations are causal explanations. Their specific form may differ, and the form of explanation in physics certainly differs from explanations in biology or psychology. Still, I believe we can explicate all of these as different forms of causal explanation, and I will go into this in a bit

more detail later (see section I.6.2). What I would like to point out for now is that, looking into the sciences, the explanatory value of a theory is determined in relation to available alternatives. For example, Newtonian physics describe a great deal of physical occurences on a mesoscopic level quite accurately, whether it be the curve of a thrown object, the acceleration of an object on which a growing force is exerted, and so on. For this reason, Newtonian Physics has been doing a stellar job and was for a while considered *the* standard for physical explanation. However, Relativity and Quantum Mechanics are far superior when it comes to describing and explaining occurences on microscopic and astronomical levels, and the occurences on mesoscopic levels which are so well described by Newtonian Physics can (with a grain of salt) be derived as special cases of General Relativity. Thus, Relativity and Quantum Mechanics can be said to explain physical occurences much better than Newtonian Physics, and with the former's advent, the latter's limitations in describing our universe became apparent. Today, Newtonian Physics is usually considered an approximation of the reality of physical laws – that is, it does not state the "real" facts of the matter, but useful approximations.

Applying this relativity of explanatory force to matters of psychology, it may be the case that Paul Ekman's theory of basic emotions (see Ekman 1999) can explain why emotional expressions are found to be quite homogenous across different cultures. However, if we had a complete theory of the evolution of emotional expressions, involving an integration of, say, migratory patterns of the human species across the planet over the past few million years, it would certainly dwarf Ekman's theory when it comes to explanatory power. That is not to say that either Newton's or Ekman's theories do not explain things to a certain degree; it's just that if there are better theories, then we are more likely to accept the ontologies of the better theories, and to view the worse theories as useful tools instead of reflections of the reality of things.

Coming back to mental states, this means that in order to find out what mental states are, we do not simply look to any explanatory theory, but to the best (or at least to a group of those which are currently considered best – there may not be a decisive criterion for which theory is actually "the" best among a group of competing theories). This notion chimes with Quine's idea that *what there is* is not decided by a-priori ruminations, intuitions or a direct non-inferential perecption of external objects, but by what objects are assumed by the best scientific theories (see Quine 1980: ch. 1). Briefly, according to Quine, the question "what is there?" – i.e. what sort of entities exist – is decided by looking at what the bound variables are in our best scientific theories. What the best theories are will change over time, but that

explains why our ontologies do not stagnate, and why today we are more prone to believing that the Higgs Boson exists rather than the Pantheon of Greek Gods, even though the latter were invoked to explain observed events: Hephaestus' activities were meant to explain volcanic activity and Zeus' wrath to explain thunderstorms. In this sense, ontology is relative to the theories we adopt, but this ontological relativity is neither arbitrary nor random, and thus cannot lead to a radical skepticism concerning the existence of the external world; for what theories we adopt is tied to what external constraints are imposed on us. We do not adopt theories arbitrarily or at random, and the mere practice of evaluating theories according to their explanatory value assumes that there is an external constraint which is not under our own control, but reflects aspects of an independent world. How exactly these external constraints interact with our theories is a different question, but I believe it is safe to say that we may well abandon all science if we seriously came to doubt this fact, and that, conversely, all those who have followed me thus far in conceding that the cognitive sciences actually deserve the monicker "sciences", and that sciences in general do exist, will also follow me in this matter and not yield to radical skepticism.

However, it is also true that there is more to our evaluation of what a good theory is than the constraints of an external world. Not only do we have matters of theoretical aesthetics to consider – such as parsimony, symmetry and the like –, which reflect our tastes and our mindsets more than the things with which the theories are concerned (except, of course, in those cases where the theories are *about* our tastes and mindsets), but also, in order for theories to be good, they must be able to explain matters, and as I made clear at the outset of this chapter, being able to explain things is an intensional notion, and so their being good should crucially depend on our mindset. While explanations in the sciences are different from semantic explanations, scientific explanations are not free of considerations of intelligibility. For example, there certainly is a difference between the explanations to "what does the word *bachelor* mean?" and "why does the sun rise every day?". Both questions can be taken to demand explanations, but the former question demands *nothing but* intelligibility, because semantic explanations are explanations meant to facilitate linguistic understanding. Whereas for scientific explanations, intelligibility is a requirement, but not their be-all and end-all. The minimum requirement is that *someone's* knowledge is sufficient to make the explanation intelligible, but not *everyone's*. I imagine all of us accept a lot of scientific explanations to actually qualify as explanations, even though we lack the knowledge to render them fully intelligible to us. However, we would demand that at least those who have the required knowledge in the respective field do understand the explanation, meaning that they could give

a comprehensive account of how and why the provided explanation actually explains the matter.

It is these questions of intelligibility which leave us with more things to consider than just objective, external constraints on our theories, even though the theories are ultimately *about* these external constraints, namely, about the external world. For example, two theories may be about the same things, but one may be more intelligible, making it superior to the other, lending it superior explanatory value. At other times, the available evidence may make it impossible for us to decide between two competing theories, even though they seem mutually unreconcilable. And thirdly, there may be a fundamental indeterminacy at work, which may make it impossible for us to decide between competing theories, no matter how much evidence is invoked. The latter has been claimed for the case of intentional theories, popularly by Quine, Davidson and Dennett.

In Davidson's case, the indeterminacy would boil down to using different "scales" in a description of intentional states. Davidson uses this comparison to make clear the holistic nature of the theory of mental ascriptions:

> "Just as we cannot intelligibly assign a length to any object unless a comprehensive theory holds of objects of that sort, we cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions. There is no assigning beliefs to a person one by one on the basis of his verbal behaviour, his choices, or other local signs no matter how plain and evident, for we make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest. It is not merely, as with the measurement of length, that each case tests a theory and depends upon it, but that the content of a propositional attitude derives from its place in the pattern" (Davidson 1980: 221).

According to this conventionalist view of semantics (cf. Field 1975), there can be indefinitely many theories of interpretation which are equally suited for interpreting the available evidence; this fact might not be apparent to us, since we are able to alternate between different similarly probable theories while trying to interpret another person's utterances. As long as it is not entirely clear to us what an agent means, several possible theories remain suspended and readily available – the fact that we eventually opt for one over another does not imply that we have decisive evidence for it, but will rather reflect a pragmatic choice, such as our need for a quickly and flexibly usable theory (cf. Davidson 2001a: 214). Since what counts as an acceptable theory is measured by how well it predicts "the truth conditions of sentences"

(Davidson 2001a: 74, see II.8.4.2), it is apparent that indefinitely many theories could in principle satisfy this criterion, just as indefinitely many numbers can express the fact that one object is three times as large as another (and here we shouldn't just think of different scales, but also all equivalent expansions of the fraction 1:3, such as 2:6, 3:9, 4:12 etc.).

Dennett, on the other hand, has a more fundamental disagreement between intentional theories in mind, which goes beyond stating them in terms of something akin to differing scales of measurement, and it allegedly makes him "less of a realist than Davidson (…). I see that there could be two different systems of belief attribution to an individual which differed substantially in what they attributed – even in yielding substantially different predictions of the individual's future behaviour – and yet where no deeper fact of the matter could establish that one was a description of the individual's real beliefs and the other not" (Dennett 1991b: 49). The issue of indeterminacy is a prominent one in theories of intentionality (see also Levine 1987: 272 f.), but for now, the point for me to make is simply to point out that what is a good theory depends on different factors, some of which can be related to external constraints, others to theoretical aesthetics, and still others to matters of intelligibility and, beyond these, it is possible that indeterminacy will still leave us with several mutually irreconcilable theories. However, the theories we end up with are nevertheless worth being called "our best theories", and they are what informs our ontologies.

I do not believe that such a picture warrants any immediate conclusions pertaining to the reality of the concepts under theoretical consideration. If anything, the theory-ladenness of psychological explanation does not rob the things denoted by its kind-terms of their reality, but adds further criteria for them to meet in order to enter the theory besides their being real: they need to yield an explanatory surplus. That is, introducing a mental state into a psychological theory implies that doing so is of explanatory value; for example, being able to ascribe a specific belief to someone implies that the belief so ascribed explains something which the ascription of a different belief (or ascribing the lack of it) would not explain. And that it explains something entails that there are objective properties of agents which are in need of explaining – thus, for a mental state to explain it, it must refer to something which has, at least in principle, aspects which go beyond what is radically subjective (see section I.7). Mental states may also have purely subjective aspects: my childhood memory of visiting Legoland in Denmark is associated with a nostalgic feeling which is only accessible to myself. But memory itself would not be a psychological term, and the memory's content unascribable, if there were no intersubjectively accessible properties which could be explained by ascribing this memory to me or to others.

I.6.2. Nomological Explanation

As I have mentioned in section I.5, I take psychological explanation to be a form of nomological explanation (cf. Goldman 2012: 403). Nomological explanation is a form of explanation which relies on causality, laws and kinds. Elevating a certain general term to the status of a scientific "kind" marks a distinction between general statements which are inductively supportable by their instantiations (i.e. "projectible", see Quine 1969: ch. 5) by picking out causal relations between kinds, and those which are not. We call the projectible statements "laws"; their lawlike status is what allows us to make predictions (looking to the future) and give explanations (looking to a past event).

To borrow an example from Nelson Goodman (see his 1983: 18 f.): The fact that all coins in his pocket are made of silver does not give us evidence to suppose that the next coin which is put into his pocket will be made of silver as well. Just the opposite: If this next coin wasn't silver to begin with, we have very good reason to believe it will in fact *not* turn silver merely by being put into Goodman's pocket. In spite of this, knowing that all pieces of butter have always melted when heated to 150°F, we would reasonably want to conclude that if the next solid piece of butter is heated to 150°F, it will melt as well. The notable difference between these two generalisations, which makes the latter a natural law and the former absurd, is that the relevant concepts (butter, change of temperature, state change from solid to fluid) either themselves refer to natural kinds or are restatable in such terms.

Nomological explanations work by establishing that a given event falls under a general law, and the fact that it does so explains the event in question. For example, the fact that a given piece of butter melts when heated to 150°F is explained by the fact that there is a general law from which follows that butter melts at 150°F. That there are circumstances under which laws fail to hold are specified by so-called *ceteris paribus clauses*. For example, water fails to freeze at 0°C if it is stirred. So, for the law to be true, one of its ceteris paribus clauses needs to exclude stirring. Laws themselves are typically explained by integration into higher-order laws: for example, the chemical properties of water can be integrated into physical laws, and one job of physicists consists of seeking out more general laws from which these more specific laws can be derived.

While these ideas about nomological explanation should be applicable widely enough, we should not assume that the term "law" always applies to the same thing across different disciplines. To pick a field particularly unlike physics, let's say that if some might want to posit laws in the theory of art, then they would possibly want to do so without relying on

anything like an inductive empirical confirmation of these laws. Others might want to hold that there are conceptual laws – laws which mathematics, logics and philosophy make use of –, which might, again, not have anything substantial to do with empirical confirmation. In such cases, what I am going to claim in the following need not apply. I will solely be concerned with sciences whose laws stand in an important relation to empirical evidence and/or can be said to govern what happens in the empirically accessible world. Whether the use of the term law actually does vary systematically is not the subject of my present investigation – I merely suggest that, given my examples above, if we were to ask scientists from different disciplines, we might get differing answers, and if we were to look at how they actually use the term, we might come up with substantially differing analyses of this use. I would not even expect to find a systematically homogeneous use of the term even within the cognitive sciences, or those sciences adjacent to psychology, the neurosciences, and philosophy of mind. All that presently matters is coming up with an analysis of this term which can be used sufficiently similarly across the relevant disciplines. Thus, the following is not meant as an empirical description of how scientists in the field use the term "law", but rather, how we can use this term (and related terms) to aptly describe what these scientists are doing.

In the cases I just restricted my analysis to, saying that a generalisation is lawlike is the same as saying that it is supported inductively by evidence about the properties of its objects, the evidence being singular empirical statements. These singular statements are typically observations of a specific event occuring or state of affairs holding at a given time. If this specific event or state of affairs is confirmed to systematically reoccur or hold across different observations, and there is an explanation for this permanence, then the viability of the explanation is said to be supported by the singular empirical statements – the evidence. Crucially, the relation between the lawlike explanation and the singular statements is not just one of summing up the singular statements. Rather, there has to be a criterion for whether particular singular statements can support generalisations. Compare, once again, Goodman's case of having but silver coins in one's pocket – no singular statement about all coins in one's pocket being silver at a given time (or several such statements at several different points in time) supports any general statement about all coins in one's pockets having to be silver at all times.

But what could this criterion be? According to Davidson, it always involves a petitio principii:

> "Lawlike statements are general statements that support counterfactual and subjunctive claims, and are supported by their instances. There is (in my view) no non-question-begging criterion of the lawlike, which is not to say there are no reasons in particular cases for a judgement. Lawlikeness is a matter of degree, which is not to deny that there may be cases beyond debate. And within limits set by the conditions of communication, there is room for much variation between individuals in the pattern of statements to which various degrees of nomologicality are assigned. In all these respects nomologicality is much like analyticity, as one might expect since both are linked to meaning" (Davidson 1980: 217 f.).

Said petitio principii consists in a tightly woven conceptual circle between the terms "law", "natural kind" and "projectibility". As just mentioned, projectibility is a general statement's property of being supportable by singular statements which count as evidence for the general statement's truth (cf. Quine 1969: ch. 5). The circle goes like this: Natural kinds are objects which are projectible, they are projectible if they figure in natural laws, natural laws are generalisations of singular statements about natural kinds. Virtuous as it may be, the analytic circle is tight, and thus, on the face of it, unsatisfying. However, satisfaction can be gained by looking at a specific example of how exactly laws get to be explanatory: Picture water – H_2O in its liquid form, between 0° and 100° Celsius – being heated to a temperature of 100° C or more, thus vaporising, undergoing a phase transition from liquid to gaseous. An appropriate lawlike generalisation would be "ceteris paribus, if H_2O is heated above 100°C, it vaporises". It is lawlike because it is projectible: If H_2O will be heated above 100°C, then, ceteris paribus, it will vaporise – tomorrow, the day after, next year, or whenever. This law is confirmed by its singular instances: Ceteris paribus, any instance of H_2O which is heated above 100°C and vaporises confirms the respective law.

While, as I pointed out, I would not dare assume that all theories in any scientific field consist of natural kinds and laws in the same sense that theories in the natural sciences do, I urge to concede that at least a set of analogous terms is available even to those sciences which do not fit squarely into the category of the "natural sciences". How and in what sense would these be analogous? For one, they are kinds, but not exactly natural. For instance, Ian Hacking, while ultimately concluding that "there is no such thing as a natural kind" (Hacking 2007: 203), sides with William Whewell (who in the mid-1800s informed the scientific use of the term "kind") in asserting that "Whewell was, in my opinion, on the right track when he said that a kind is a class denoted by a common name about which there is the possibility of

general, intelligible and consistent, and probably true assertions" (ibid.: 238).[26] That Hacking can deny the existence of natural kinds, all the while asserting that kinds do exist (namely in the form of a certain class), has to do with his taking "natural" as a concept heavy on irredeemable metaphysics. Consequently, he aims to rid the notion of kinds of the notion of having to be "natural". Hacking himself is especially concerned with kinds of humans (such as those marked by a certain mental disorder), and thus it makes sense for him to introduce the distinction between *interactive kinds* and *indifferent kinds*: Kinds of humans qualify as interactive kinds since

> "people are agents, they act, as the philosophers say, under descriptions. The courses of action they choose, and indeed their ways of being, are by no means independent of the available descriptions under which they may act. (...) What was known about people of a kind may become false because people of that kind have changed in virtue of how they have been classified, what they believe about themselves, or because of how they have been treated as so classified. There is a looping effect. (…) [On the other hand, q]uarks are not aware. A few of them may be affected by what people do to them in accelerators. Our knowledge about quarks affects quarks, but not because they become aware of what we know, and act accordingly" (Hacking 1999: 103 ff.).

Which is why quarks qualify as indifferent kinds. Similarly, Kusch suggests a distinction between social, artificial and natural kinds (Kusch 1999: 257) in place of the familiar monolithic concept of natural kinds.

It is easy to see that the natural sciences, insofar as the properties of the objects they are researching are not determined socially or agentially, could afford to rely on a monolithic notion of "natural kinds". As Hacking has pointed out, quarks are indifferent to our social practices, and proper results about them are free of the determinants of human intervention.[27] However, since we are wading in murkier waters, and the cognitive sciences are by definition pervaded by interactive kinds, we would do well not to blindly accept said monolithic notion.

---

[26] For on overview of the historical background, focussing especially on the debate about natural kinds between Whewell and John Stuart Mill, see Snyder 2006: chapter 3.

[27] Of course, observation can already be construed as intervention, especially when dealing with quantum effects. However, the mere fact that an object exhibits an effect caused by its observation alone does not preclude the possibility of classifying these effects as "natural" and the laws pertaining to these effects as "natural laws". Going back to Hacking's quote, quarks are not aware; that is, even if observation has an effect on them, these effects cannot be explained by their being agents or by their being social (or generally, by their being anything but natural). Thus, the salvageable distinction is that between effects exhibited by agents and those exhibited by indifferent objects (where "indifferent" does not mean "not being affected by observation" but rather "not reacting to observation under agential descriptions or explanations"). (Compare Dennett in Edmonds & Warburton 2015: 130 and section I.4.4.)

Getting rid of it can only mean a gain in methodological accuracy, and not at all a loss in objectivity or scientific standards: For example, in case of the neurosciences, it merely means acknowleding that parts of human brains, whose causal properties we seek to specify (thus making them "kinds" in neuroscientific theories), change depending on social interaction, and that some of these interactions consist in what we subsume under "scientific practice". That is, not only does something in an experimental subject's brain change when performing a given experimental task, and not only does a neuroscientist's brain change whenever she performs neuroscience, but theories in the cognitive sciences in general affect the properties of those kinds which figure in such theories. When concerned with investigating the human mind, overlooking this interactivity would distort results.

I take the concept "kind" to be the basic notion about which Hacking says that it denotes said class "about which there is the possibility of general, intelligible and consistent, and probably true assertions", and I will restrict my use of it to scientific contexts. Since I am not going to say anything crucial about sciences which have no laws at all, but will stay within the domain of the cognitive sciences, which I assume has laws, I will further restrict it to the use in laws. That is, kinds are what scientific lawlike generalisations can be made about. My take on what is properly scientific is rather lenient, insofar as I admit not only laws which are supported inductively and empirically, or laws which can be stated in quantifiable terms, as properly scientific. Rather, I admit both quantitative and qualitative laws, and I admit many different forms of systematic generalisations which yield explanatory surplus in the cognitive sciences, even though some of these have a reputation of not being completely empirical. For instance, many philosophers and psychologists suspect that psychological laws are neither as strict as physical laws, nor could be completely rid of intentional and semantic terminology. This is what I am going to elaborate on in the following sections.

I.6.3. Intentional Explanation

Explanation by intentional mental states is the form of psychological explanation which has classically been of special interest to the analytic philosophy of mind. Yet, its use to cognitive science remains controversial. Between Jerry Fodor, who believes there is no serious rival to its explanatory power (cf. Fodor 1989: 6), and Patricia and Paul Churchland, who endorse the eventual abandonment and replacement of the "propositional attitudes" by neuronal states (see e.g. Churchland 1981), virtually every possible position can be placed. I

am not going to discuss eliminativism here, except for the brief suggestion that, if we are to construe it as an unshakable faith in a future abandonment of mental states, this position should appear as a risky gamble, depending on "presumptive theses way out in front of the empirical support they require" (Dennett 1991b: 51). It is far from clear how current research could support the theoretical elimination of mental states (cf. Gold & Stoljar 1999).

While I aim to make a stronger case for the explanatory value of intentional psychology, a minimal case can be made for the importance of taking intentional explanation seriously for the sake of interdisciplinary communication:

> "At a bare minimum, trying to understand the relationship between the *intentional stance*, which common folk and some scientists take towards human organisms, and the *physical stance*—the assumption that an organism's behavior has internal physical (e.g., neural, biochemical) causes—seems prerequisite for effective interdisciplinary communication. No reasons exist to think practitioners in areas of science outside of neuroscience will completely abandon their appeals to folk psychological explanations of behavior, nor is the eliminative materialism for which Paul Churchland (1981) advocates obviously in the offing. Furthermore, given that misunderstandings between neuroscientists and ordinary folk who are looking towards neuroscience for answers may also arise, it seems legitimate for the sake of clarity for neuroscientists to be clear about how they understand the mind and how and in what ways that differs from how non-scientists think about it" (Sullivan 2014: FN 5, 63f.).

Now, how do intentional states actually explain behaviour? For illustrative purporses, consider the following example: Suppose Kate and Henry are invited to a social event, and suppose also that they know they are required to bring food and beverages. Since it is more convenient for them to split these tasks, Kate expresses her intention to Henry to bring food, knowing that Henry will then bring beverages. If we witness Henry's buying the required beverages in time for the social event, we will explain this by mentioning some crucial part of this story. And if we are to get to the bottom of the explanatory role of mental states, we ought to give a full, non-elliptic explanation and see what it consists in. Thus, we are required to make Kate's knowledge that, by making explicit to Henry her intention to bring food, she intends to persuade the latter that he'd best bring beverages a necessary part of it. The fact that this ascription of this very mental state to Kate is essential does not merely follow from Henry's buying beverages in time for the social event, since Henry might very well have not been listening to Kate at all; he could have been distracted and not register Kate's assertion that she intends to buy beverages, and have had an independent reason to buy them, thus

rendering Kate's mental state explanatorily inert. But that is not what happened, and our requiring a full explanation for this case is to request more than giving an explanation of just *any* instance of Henry's buying beverages: It is to request the explanation for *this* special case just as it happened at this point in time, given *this* sequence of events. And in this case, we know that Henry's being persuaded by Kate led to his buying beverages (rather than just, say, his knowing that beverages had to be bought by someone in time for the social event). This insistence that only certain salient features of the environment actually explain an action, as opposed to those features of the environment which *could* reasonably prompt it, hints at the need for accounting for epistemic properties of the subject whose actions are to be explained: Only those features of the environment which reasonably prompt an action and which are cognitively available to the agent can count as explanatory.[28] The fact that Kate's persuasion was effective in this case, but might not have been effective in other cases in which Henry bought beverages for other reasons, can be highlighted by saying that Kate's persuasion was the *cause* of Henry's buying beverages (here, I am following Davidson's highly influential account of causal action explanations – see Davidson 1980: 3-18).

So, Kate's mental state is essential for explaining Henry's action because there is a causal connection leading from the former to the latter. How does this connection come to pass – i.e. what are the relevant parts of the underlying mechanism? Firstly, it critically involves some properties of Kate's which led Henry to believe that she promised to bring food. These properties have to fulfill two requirements: They have to be expressions of Kate's mental state, and they have to be observable (or, more specifically: in order to be of explanatory value, a sufficient amount of these have to actually be perceived by Henry). That they have to be expressions of Kate's mental state is to say that Henry's interpreting these as being evidence for Kate's mental state is justified. Henry could be so confused as to interpret any perceivable set of properties of Kate's as expressing any arbitrarily assigned mental state; but in most cases, he would be objectively wrong, and he would be wrong according to intersubjectively available justification conditions for the ascription of mental states (see section I.7). If there were no such conditions, no one could ever be wrong in their ascribing any mental state to any person.[29] What he needs to manage is to connect Kate's observable

[28] This explains why intentional ascriptions create *intensional contexts*: Contents in intentional ascriptions can only be substituted by those which have the same extension *and* the agent knows about (see Quine 1980: ch. 8). When judging whether someone knows that Reginald Kenneth Dwight has been knighted it does not only matter that in fact he and Elton John are the same person and that she knows that Sir Elton John has in fact been knighted – what also matters is whether she knows that "Reginald Kenneth Dwight" and "Elton John" designate the same person. See section I.6.1.

[29] Assuming that the practice of ascribing mental states is not a big hoax, I will take my following elaboration of an account of what these conditions consist in to be more worthwhile than their justification. Of course, if there is no independent, empirical proof that mental states have explanatory value, then this account will be circular:

properties systematically to his ascription of mental states in a consistent way; among other things, that means that he himself can explain Kate's actions by referring to the mental states he ascribed to her. For instance, his interpreting Kate's observable properties as meaning that she loves ice cream should put Henry in a position to be able to explain her grabbing more ice cream than, say, Bob (whose observable properties allow Henry to ascribe to him the mental state that he's indifferent toward ice cream).

The observable properties which justify the ascription of mental states such as "liking ice cream" will obviously have to go beyond post-hoc ascriptions. For instance, in order for the mental ascription to be a predictor worthy of its name, the ascription should be in place well before the bearer of the mental state in question actually goes for the ice cream in any event which is to be predicted. Also, for mental states to exceed purely behaviourist notions, they should be more than dispositions to behaviour, and to some degree theoretically independent of the actual behaviour associated with them. While behaviour is constitutive (or "criterial") for ascribing mental states, explanation by mental states is not behaviourist explanation (see section I.7.3): psychological states need not "draw inferences from behavioral evidence, [but] (…) the fact that overt behavior *is* evidence for (…) [them] *is built into the very logic of these concepts*, just as the fact that observable behavior of gases is evidence for molecular [states] (…) is built into the very logic of molecule talk" (Sellars 1997: 107, §59).

It would also be mistaken to insist that behaviour being constitutive for mental states implies that they can only be had if they result in behavior. For example, a patient with locked-in syndrome can plausibly have mental states. She could be able to think and feel without any of us noticing, and without any of her mental states ever resulting in overt behaviour. What the grounding in behaviour is meant to imply is that the connection between mental states and behavior cannot be theoretically severed: That is, mental states are kind-terms which exist because the theories which give them their meaning explain behavior. Whenever a patient with locked-in syndrome has a mental state, she has a state which is paradigmatically invoked to explain behavior, a state whose theoretical significance lies in explaining behavior. Still, it can practically occur without doing its explanatory job, and it just

---

for then it will look as though I'm assuming that mental states have explanatory value because there is a practice of ascribing mental states, and that there is a practice of ascription because mental states are of explanatory value. Therefore, I claim that such proof can be given independently, and I imagine this proof to proceed along the lines of Fodor's defense of the propositional attitudes (in his 1989: ch. 1). It consists in facts like these: If A intends to be at a conference next Tuesday, then knowing his intention is a better predictor for A's whereabouts next Tuesday than any other (non-mental) fact about A.

so happens to be unable to explain any behaviour of a patient who is physically kept from behaving.[30]

Intentional explanation as just illustrated is a form of causal explanation, and it works by stating the relevant beliefs, desires, intentions or other propositional attitudes which have caused the explanandum. Even though, as I have made clear in section I.2, intentional and intended are not synonymous (i.e. intentions are a subclass of intentional states), actions are both intentional and intended; they are intentional insofar as they are aimed at something (and explained by mental states whose intentional objects are explanatorily related to the action in question), and they are performed with an intention. The general form of this explanation (following Davidson 1980: 5) is this:

(I) A desires to bring about X
(II) A believes that doing Y leads to X
(C) A does Y (or at least intends or is motivated to do Y).

Here, Y is the explained action, and it is explained by making explicit A's desire aimed at X and her belief that Y leads to X. A's doing Y is being explained by stating (I) and (II) because it is specified by the conclusion following logically from (I) and (II).

Still, the logical form alone is not the whole story when it comes to the explanatory force of intentional explanations. Rather, the logical form above describes what we are prone to accept as intentional explanations. But why do we accept this form of explanation in the first place? Or, more specifically: Why do explanations of this form explain anything in the sense of making the respective action *intelligible* to us?

This is because intentional explanation is systematically intertwined with intelligibility; namely, with meaning and understanding. Following Davidson, understanding meaning is a matter of attributing rational intentional mental states:

> "we could not begin to decode a man's sayings if we could not make out his attitudes towards his sentences, such as holding, wishing, or wanting them to be true. Beginning from these attitudes, we must work out a theory of what he means, thus simultaneously giving content to his attitudes and to his words. In our need to make him make sense, we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying)" (Davidson 1980: 222). "Any effort at increasing the accuracy and power

[30] Compare the analogous case Block makes for patients with locked-in syndrome potentially being conscious, while being entirely unable to *report* their being conscious (Block 2007: 483 f.).

> of a theory of behaviour forces us to bring more and more of the whole system of the agent's beliefs and motives directly into account. But in inferring this system from the evidence, we necessarily impose conditions of coherence, rationality, and consistency" (ibid.: 231, see also 241).

That is, by attributing intentional mental states to someone, we understand them, and their function of explaining actions is the very basis for attributing intentional states to someone. Since actions are intentional to begin with, they are in specific ways related to the environment by being directed at certain parts of it. Thus, the agents' specific directedness at their environment goes hand in hand with what they believe to be the case in this environment, what they desire from it, and so on. While there are shortcuts to attributing mental states, such shortcuts fundamentally depend on a (as Davidson says, *by our own lights*) rational and consistent connection between assigned intentional mental states and observed actions in a meaningful environment (for the details, see section I.7.4):

Usually, we do not attribute mental states to our peers from scratch – that is, we do not have to invoke the ultimate bases of mental states in order to attribute them. Rather, we simply go ahead and assume a lot of these states based on contextual cues, such as social context, self-reports and/or third-person reports. Often, the fact that someone is a bureaucrat alone explains a lot of their actions directly, since we simply assume a great deal of their intentional states (and if we first observe someone's actions without knowing they're bureaucrats, being told that in fact they *are* bureaucrats can explain a lot as well). And being told that someone is a party member explains why they raise their hands during voting at their party rally, in a way that being told that they are a bystander does not (compare Danto 1973: ix f.).

I.6.4. The Normativity of Intentional Explanation

As we have just seen, psychological natural kinds may seem a bit odd when compared to those invoked by natural sciences like physics or chemistry, since they are related to one another not only by causal laws but by logical and normative ones as well:

> "[I]t is the myth of our rational agenthood that structures and organizes our attributions of belief and desire to others and that regulates our own deliberations and investigations. (…) Folk psychology, then, is idealized in that it produces its predictions and explanations by

> calculating in a normative system; it predicts what we will believe, desire, and do, by determining what we ought to believe, desire, and do" (Dennett 1987: 52).

For instance, if I promise to give a speech at a conference next month, then that promise (together with the assumption that I understand what promises are, and that my psychological constitution is such that I usually keep them) supports the prediction that I will be at that conference next month. Why? Because I had better! And if I believe that keeping promises is a good thing, and being at the conference next month is a way of keeping my promise, then that also supports the prediction that I will be at that conference next month. Why? Because it's only logical! There is little doubt that by way of their predictive and explanatory power, normative and logical relations such as the ones invoked in these examples figure in psychological laws, and that when it comes to predicting and explaining human actions, these psychological laws are superior to any other laws from any other field (such as mechanics; see Fodor 1989: 6). The latter claim is vindicated by facts such as that David Cameron's often heading to 10 Downing Street after his day at work is best explained by the conjunction of his beliefs that the United Kingdom's prime ministers resides there and that he himself currently holds this office.

Given that we usually expect causal theories to not rely on normative notions – shouldn't causes rather pick out things *descriptively*? –, we need to reconcile the notions of nomological causality and normativity inherent in intentional psychology. First, let me loosely invoke some minimal criteria for what counts as a causal relationship. I am going to briefly sketch how intentional psychology meets them. But more importantly, I will show how its normative aspect is in fact conducive to its meeting them.

Firstly, we should require that A constitutes a cause of B if A brings about B. Secondly, A and B need to be types of events, not tokens. That is, singular events can only ever counts as instantiating lawlike causal relationships if these laws apply to such events in virtue of certain generalisable properties exhibited during these events, namely kinds. So, in order for beliefs, desires and intentions to have explanatory power and to constitute kinds, many different persons across many different situations need to be able to have beliefs, desires, and intentions (see I.6.5). Thirdly, there needs to be a theory from which it follows that A causes B. (This criterion takes care of our wanting to support predictions and counterfactual reasoning about causal relationships, such as "if A would have occurred, B would have occured".) Fourth, this theory should not be contradicted by a clearly superior theory. A theory is explanatorily superior to another if it explains more phenomena or if it

explains the same amount of phenomena more efficiently (by e.g. being more sparse, more consistent, better integratable into other theories we accept, or better understandable).

What counts as A's bringing about B is related to what we scientifically know about the world. That is, not only does fiction or ideological dogma not establish what causes are, but neither does much of what we had thought of as science or explanatory models in the past. Magic is not a cause of sickness, phlogiston is not a cause of combustion, and male masturbation is not a cause of depletion of the masturbator's spinal fluid. This also means that, since psychology introduces its natural kinds and its notion of nomological causal relatedness between them to explain certain phenomena, psychology's alleged causes cease to be real if a supreme science comes along, which robs psychology of its dominant status when it comes to explaining, say, why people often show up at the places they intend to show up (compare I.5). Only then will mental states turn out not to be kinds and not to be causally efficacious, in whatever sense the new science requires them to neither be kinds nor efficacious. Any science has to face this danger, and realistically, what we're going to have to deal with in the foreseeable future is not a paradigm change in the wake of the advent of a supreme science, but new psychological theories which outdate the old ones. Revisions are ever ongoing.

According to the invoked criteria, either of the previously invoked normative or logical relations, which form the basis of intentional laws, count as specifying causes, since both the norm that I should keep my promise, as well as the logical syllogism constituted by my desire to keep promises on the one hand, and my belief that attending the conference is a way to keep it on the other hand, brings about my attending the conference (or stands in some other nomological relation to my attending the conference, such as a probabilistic one). These laws are broadly applicable in virtue of intentional attitudes being kind-terms which can be instantiated across many different individuals and situations, and they are instantiated according to fixed, if usually implicit, criteria for what counts as having one such attitude (see section I.7.1.). The laws themselves follow either from the conceptual relations holding between the kind-terms (such as, ceteris paribus, desiring to drink causes drinking), from the logical form of action explanation (inherent in said syllogism, see section I.6.3.), and/or from rational norms or the psychological efficacy of reasons.

But does the normativity of intentional laws dilute the scientific quality of intentional psychology? That is, does the latter have to invoke something which is completely removed from descriptive or natural facts? Does intentional psychology amount to a mythological narrative or a form of hermeneutics rather than a "hard" objective science? To allay such

worries, it should be stressed that normative and logical relations do not *directly* enter into predictions or explanations of intentional psychology, and that there is a dichotomy between describing psychological causes and what is generally logical or reasonable in a normative sense. On the one hand, in order for a reason-explanation to work, it has to pick out a reason that is or was in fact efficacious in the agent's mind – one that was cognitively transparent to her and which caused her action. But on the other hand, in order for this reason to be explanatory, it has to be applicable in a psychological law, and thus generalisable. And which psychological law is generalisable at least partly relies on what's rational, and thus on more than the cognitive makeup of individual agents. That is, laws are often generalised because they are rational: individual agents are generally trained to shape their thoughts, desires and intentions according to what is deemed rational (both in a minimal logical as well as in a more substantial moral sense) and so instantiations of psychological laws are partly due to the respective law's being rational. Thus, external norms can be cited as causes, but not merely by being "reasonable" in an abstract sense, but by therefore being efficacious in the event that is to be explained, namely by causing an agent to share this norm and act or reason in accordance with it. Rationality can in this way shape our cognitive apparatus and therefore needs to be *descriptively* (not normatively) invoked as a cause for the structure of our cognitive make-up (see section II.3).

That is: Yes, we should accept that there *are* in fact norms, and that they potentially come from a not entirely scientific place, such as a social convention (e.g. about promise-giving and -keeping). However, some norms which intentional cognitive capacities rely on are even less conspicuous: in chapter II, my analysis will rely on evolutionary aims, which can be explicated as indicating, say, that a toad *should* catch worms (see also section I.8.4). This "norm", of course, is akin to a natural fact (given the toad's organismic structure and evolutionary history).

Some confusion is caused by the commonly ambivalent use of the term "reason", namely as referring to an actual psychological cause as well as to something which normatively governs psychological causes, both internally ("she reasoned that she should stop smoking") as well as externally ("reason demands that she should stop smoking"). To clearly bring out this distinction, we can pick out an unreasonable (i.e. normatively or logically unsound) desire as being the reason (i.e. psychological cause) for someone's action.[31] This *reason to act* is a descriptive notion when it comes to explaining action; it refers to something

[31] Classically, human psychology has often been marked using a dichotomy between reason and emotion (or desires/passions). This dichotomy is not to be transferred to models of intentional explanation, since emotions/desires/passions also constitute intentional reasons (i.e. psychological causes) to act.

which descriptively persists, namely an intentionally characterised cause for an action, something which is part of the real psychological make-up of an agent. On the other hand, what has been called reason in a second sense, namely as a faculty governing mental processes in accordance with norms of rationality, shapes our actions insofar as we can strive to act in accordance with it. For example, perhaps there is someone who desires apples more than oranges, and oranges more than bananas, but bananas more than apples. Any such preference ordering A > B > C > A is irrational in the sense that it makes us exploitable: on a behavioural interpretation, it means that we are willing to trade A and some sum for B, then to trade B and some sum for C, then C and some sum for A, ad infinitum, thus losing everything while never gaining anything – anything but the satisfaction of our irrational desire, perhaps (cf. Ramsey 1931: 156-198 and Davidson et al. 1955). Yet, while they are in this sense irrational, it may be true of anyone that they have these desires, and in such a case these could causally explain why such a person keeps losing money. Thus, an irrational desire can explain my actions, and in this sense constitute a psychological cause. On the other hand, what rationally justifies an action need not specify what actually causes an action: What we usually call "reason" is a normative ideal to guide our actions, while not necessarily constituting our actual psychological causes. Certainly, neither our mental states nor our behaviour strictly adhere to what is logical, which is why we usually don't explain actions by merely assuming that carrying them out was logical (the fact that Mr. Spock, a prime example of someone who would do this, hails from science fiction should drive this point home). However, once we are educated about the fact that an irrational preference ordering such as A > B > C > A makes us exploitable, we could be motivated to get our act together and rid ourselves of such irrational desires.

Many desires are not formally or logically irrational but practically, such as desires aimed at excessive consumption – smoking, binge-drinking, drug-abuse, and generally addiction. There is no formal argument to be made against these, but such an argument rather consists in pointing out that they have unacceptable practical consequences. Thus, we say that for practical reasons it is irrational to act on desires born out of addiction, and we should do what we can to not act on these desires (such as to seek external help). Of course, addiction can still be a descriptive reason for an addict to act, in the sense of constituting a psychological cause.

I.6.5. The Generality of Intentional Laws

The explanatory value of intentional psychology hinges on two things: How general we can expect the psychological properties specified by intentional laws to be, and what the necessity of restricting their applicability through ceteris paribus clauses ("other things being equal") implies for their explanatory value. The first point will be dealt with in this subsection, the second in the following one.

The extent of generalisability of specific psychological laws is delineated by psychological research. Some mental properties will turn out to be stable across individuals and populations, some will allow for regularities in terms of systems of classification (such as personality-types which differ between individuals but are, as a type of classification, stable across different populations), and so on. The details of such research, however, are beyond the reach of this book. Instead, what I will be concerned with in this section is rejecting the notion that psychological properties could be so spurious as to have catastrophic consequences for formulating psychological laws.

What the notion of a psychological law shares with natural laws is their generalisability in the form of F → G ("if F happens, then (necessarily) G happens", or "something's having the property F causes it to have the property G"). Any law is not merely an enumeration of instances in which F leads (or has led) to G (see I.6.2). Rather, it is the hypothesised causal link between F and G which makes us say that each instance of F leading to G supports our theory of which it is an integral part. And F explains G because F causes G. However, psychological laws differ from natural laws when it comes to the reliability of their predictions: For example, the fact that massive bridges can be built, that train tracks endure under the great stress of high-speed trains rushing over them, or that DVDs can be mass-produced and used in many millions of homes is owed to the reliability of the laws of physics. Psychological laws seem comparatively spurious: Being subjected to the same environment will still have two different people thinking different things and acting in different ways. If you'd place me in Times Square next to a stranger, chances are that our behaviour would diverge – and that our thoughts would do so even more, to the point of diverging completely. I could be thinking about how to continue writing this chapter, while the stranger next to me could be thinking about the Broadway play he's about to buy tickets for (which I'm clueless about). There are two strategies of dealing with this apparent divergence: We could explain it in terms of further, more detailed information, such as the stranger's interest in Broadway plays, which contrasts with my cluelessness about them. Since human psychology is

(evolutionarily and socially) made to be workable in everyday life, we should not expect it to yield predictions which are accurate to the n-th degree, but rather to supply us with adequate information about our (evolutionarily and socially) relevant environment, given supplemental strategies readily available to us. One such strategy is to simply ask either of us about our thoughts, and to supply our reasons for acting divergently. Far from having to depend on these self-reports as infallible sources of (or direct access to) someone's psychological make-up, they constitute one of several forms of evidence for the ascription of mental states, and should be weighed in light of all of it (cf. Davidson 2001b: 3-14). If, say, my mental preoccupation with writing this book explains my current disinterest in visiting Broadway, then chances are that I can tell you about it, or that you can at least infer my mental preoccupation from knowing about my writing this book. This, in turn, depends on whether the psychological law that people who are invested in some extensive endeavor on a day-to-day basis are likely to be mentally preoccupied with it is true.

What I have said about intentional psychology so far already implies that we should not expect our constitution to be so intricate that the information we have to gather as input for making accurate predictions about our behaviour is so specific that it cannot be generalised enough to constitute a law. That divergences between psychological constitutions cannot be so systematic and so great as to make the whole theory break down is partly owed to the fact that intentional psychology depends on learnable norms governing how to deal with representations. That is, if any intentional state can be traced back to a representational cause, such as learning a language can be traced back to some fixed rules of grammar and vocabulary, then the associated behaviour should be regular enough. For example, if any of us would only form meaningful sentences half of the time when we are expected to, then we would not call this person a competent language-user, and we would keep the attribution of the associated meanings from them, and thus the respective intentional states. Any such regular behaviour depends on underlying cognitive dispositions to learn languages, so from the mere fact that language-use is a regular phenomenon we can infer that the mechanisms enabling us to have these dispositions must have developed in a stable and general enough way to allow competent language use. And similar conclusions can be drawn in competencies analogous to language-use, such as our use of tools, the competent interaction with our peers and our environment, and so on.

On the other hand, regularity of psychological constitution also goes beyond representational causes: While it is true that there may be innumerable mental differences between any two persons, our behaviour in situations which are highly relevant (i.e. in

situations in which we depend on predicting a sizable group's individual actions, or in which we want to make sure that behavioural homogeneity and homoeostasis obtains) can be predicted and actually socially controlled to such a degree that it enables us to have institutions such as governments, medical supply, universities and the like. That is, in each instance in which behavioural uniformity obtains, the underlying psychological laws must also obtain. Divergence has its limit.

Open questions remain in domains in which social control and psychological predictability is pursued, but where it is unclear to which degree it can be accomplished. For instance, we know that google or facebook can make significant predictions about a person's future biography based on statistical inferences from past biographical information. This can be information which may seem rather obvious. For example, predicting that I will graduate with a significant probability follows already from my pursuing a PhD degree. In such cases, the issue is not whether there can be an algorithm for inferring one from the other, but rather that the necessary input is available to companies running social networks or similar internet services and that they can exploit the according output. That is, it might well be the case that some non-public information about anyone is available to social networks because these have (1) the algorithm which outputs this information based on the input of public information (and they might well be the only ones able to develop this algorithm because no one else has a statistical basis large enough to verify whether the algorithm is reliable), and they also have (2) the information needed for running the algorithm on a given individual's data – for example, it has been claimed by facebook that information about a shared circle of friends between two romantically linked persons allows them to make statistical inferences to the duration of their relationship.[32] What supports the statistical inferences need not exclusively be psychological laws, but also social commitments or peer pressure; and often, one works by employing the other. For example, as I have said, it is likely that I will graduate based on the mere fact that I am pursuing a PhD degree, which is to say little more than that I am enrolled at an institute which steers me toward this degree. But the fact that I am part of this institute is based on some of my interests, just as well as my making it to the end of the program is based on my long-term motivation, work ethic, resilience, and so forth. In all these cases, we should expect some robust psychological laws to emerge from and support biographical facts. However, since it is predominantly economic and security concerns, and not scientific

[32] Facebook data scientist Bogdan State published his results under the title "Flings or Lifetimes? The Duration of Facebook Relationships" (www.facebook.com/data, or www.facebook.com/notes/facebook-data-science/flings-or-lifetimes-the-duration-of-facebook-relationships/10152060513428859).

interests, which direct data mining scrutiny, the large-scale experiment of gathering our social and biographical data through the internet, for the time being, remains one-sided.

I.6.6. The Relative Strictness of Intentional Laws

Searle points out that "human behaviour, where rational, functions on the basis of reasons, but the reasons explain the behaviour only if the relation between the reason and the behaviour is both logical and causal. Explanations of rational human behaviour thus essentially employ the apparatus of intentional causation" (Searle 2000: 106). He characterises intentional causation as a form of causation in which "the cause and effect work in the way they do because either the cause is a representation of the effect or the effect is a representation of the cause" (ibid.: 105). For example, if my wanting to drink a glass of water does cause my drinking a glass of water, then it is my attitude (namely my desire) toward drinking a glass of water which has brought about my drinking. Thus, the cause was a representation of the effect. Similarly, if I correctly remember that Daniel Day-Lewis won three academy awards, then this mental representation was caused by Daniel Day-Lewis's actually winning three academy awards, making the effect a representation of the cause.

Since actions are a subclass of behaviour, namely that which is caused by reasons, we can simply refer to what Searle calls "rational human behaviour" as actions. As an illustration, compare a knee-jerk reaction to kicking a drum. If the former is caused by a doctor's striking her patient's patellar ligament with a reflex hammer, then the movement of the knee has not been intended by the patient, and should be considered unintended behaviour, just as we would a nervous twitch or a stammer (at least assuming that these are not intentionally performed, say, by Dustin Hoffman playing Raymond Babbitt). But if the latter is caused by a drummer's wanting to test his equipment, his knee movement is very much intended and thus qualifies as an action. As pointed out by Searle and in sections I.6.3 and I.6.4, action explanations work the way they do because the action is both rationally or logically derivable from those reasons given in the explanations as well as caused by these: That is, if I just drank a glass of water, then your knowing that I was thirsty and that I believed I could quench my thirst by drinking a glass of water explains my drinking (compare Davidson 1980: 3-18). However, logical relations like the one supporting the syllogism "Drinking a glass of water is a way to quench thirst; I am thirsty; thus it would be reasonable for me to drink a glass of water" do not, by themselves, establish that the conclusion is actually caused by the two

premises. Because even if it were reasonable for me to drink a glass of water, two things may keep me from drinking: reasons speaking against drinking, and any type of external (i.e. non-mental) obstacle. We distinguish these two cases by saying that we either decided against drinking or that we were kept from drinking. The first implies rational control of the agent, the other a non-mental obstacle (which may be a brick wall just as much as a disease – meaning this obstacle does not have to be external to the body, but rather beyond agential control). It is the first case with which we are concerned in intentional psychology, since it says something about the agent's mind where the latter one does not: Because even if a thirsty person eventually decides against drinking, the fact that she had a reason for drinking is not to be disregarded in an account of her mental state. Rather, we say that the agent had conflicting reasons, and if we wish to continue ascribing rationality to her, we should want to say that the stronger reason won out and caused her not to drink.[33] Thus, the form of a causal explanation is maintained, even if some reasons (such as her thirst) ultimately proved not to be causally effective.

In a nomological account of mental states, the relation between thirst and drinking only holds *ceteris paribus* (which translates to "other things being equal"). That is, when an agent is free to drink, then, in the absence of stronger reasons speaking against drinking, thirst causes her drinking (compare Nachev & Hacker 2014: 200). A second precondition for such a relation to hold is that the law's specified consequence is under agential control (in our example: that the agent is actually free to drink by not being restrained, by being able to reach the glass, etc.).

There has been some debate about whether psychological laws differ substantially from other laws, especially from those in the natural sciences, when it comes to the aspect of ceteris paribus clauses (cf. Boyd 1999). For example, Donald Davidson held that it is necessary to an account of the nature of physics that physical ceteris paribus laws are required to be translatable into (or reducible to) accounts which contain no ceteris paribus clause at all (I say "accounts" because even though we may still call them laws, they may formally be very different from the laws we are used to). For example, if we wished to determine tomorrow's movement of Mars relative to our solar system, we would in this account include a finite amount of forces which act on Mars, primarily the other celestial bodies in our solar system. However, to accurately determine Mars's movement, a complete account of the state of the whole universe would have to enter into our calculations. Not only are additional forces

[33] Sometimes, weaker reasons may win out, in which case we may speak of akrasia or weakness of the will, and deem the agent irrational, effectively revoking agential ascriptions (compare Davidson 1980: 21-42) – but I will not pursue questions concerning weakness of the will here.

exerted on Mars by celestial bodies external to our solar system, but there may also be asteroids entering our solar system tomorrow, thus be "internal" to our solar system when they are still "external" today. We would in fact have to assume that all relevant bodies and forces will act on Mars tomorrow as we actually expect them to act (ruling out the sun's going nova, a radical increase in its gravity, and so on). All of these assumptions amount to ceteris-paribus-clauses.

However, "Mars" is actually not a physical description at all: Mars is not a natural kind in physics, but in astronomy. So, questions about the movement of Mars may not even be coherently posable in an ideal physics. Rather, we would ask about the location of all physical particles which make up Mars (and what "makes up Mars" is itself not a matter of physics). The same goes for the non-physical terms "solar system" and "tomorrow". Ideally, physics would describe an interaction of "ultimate" physical particles, and as Davidson believed, it would not even describe these interactions in terms of causes and effects, since the description of ideal physics would be one of a totality of circumstances.[34]

While we run into similar complications with interdisciplinary reformulations of mental kinds, it seems plausible that, if Davidson was correct and physics can get rid of ceteris paribus clauses, there is in fact a crucial difference to psychology: There seems to be no single psychological law which holds under all circumstances, even if we were to consider the totality of all mental phenomena. For one, all psychological laws which are concerned with internal cognitive states must at least assume that the agent's brain doesn't short out as a consequence of the relevant cause, never bringing about the mental effect specified by the law. Still, that doesn't diminish the explanatory value of psychological laws, since firstly, the fact that psychological explanation needs to be supported by ceteris paribus clauses meant to exclude *non-mental* interferences does not imply incomplete *mental* explanation. (But this just means that psychology is not a universal science: certainly not every event or phenomenon which does, will or can manifest itself is a *psychological* event or phenomenon.) Rather, that it is possible for a mental effect to not be brought about even when its nomologically specified cause holds (e.g. because some necessary neural connections break down between the causal effect of a mental law's antecedent and the obtaining of its consequence) does not diminish mental explanations, since the relevant explanation is the subject of non-mental

[34] Davidson takes causes to be singled out from this totality, in the sense that they are interest-relative or selective explanations. "Explanation in terms of the ultimate physics, though it answers to various interests, is not interest relative: it treats everything without exception as a cause of an event if it lies within physical reach (falls within the light cone leading to the effect)" (Davidson 2004: 113), whereas "mental concepts (…) appeal to causality because they are designed, like the concept of causality itself, to single out from the totality of circumstances which conspire to cause a given event just those factors that satisfy some particular explanatory interest" (Davidson 2001b: 216).

explanation (for example, why the neural connections broke down).[35] Inversely, natural laws can fail to hold for mental reasons: A bowl of water can fail to freeze at 0° C because I had the desire to stir it shortly before it hit 0° C. That does not make the generalisation that water freezes at 0° C any less of a law.[36] And even if physics turns out to be complete and psychology fundamentally incomplete, that still puts psychology in the same group as, say, biology – a science which certainly does not explain everything, but which has no serious rival in its proper domain.

Secondly, all *mental* ceteris paribus clauses are explanatorily valuable and potentially lawlike themselves. That is, my desire to drink a glass of water can fail to cause my drinking a glass of water because another person also desired to drink it – which she did, thus keeping me from doing the same. Or my thirst failed to make me drink because I was in a rush and had no time to do so. In which case it is assumed that, given the circumstances, I found it reasonable not to drink, again maintaining the form of action explanation by citing mental reasons which are (at least potentially) lawlike themselves.

Thus, while conceding to Davidson that psychological laws cannot get rid of ceteris paribus clauses, we should not jump to the conclusion that this fact fundamentally diminishes their explanatory value; it simply highlights one limitation. For Davidson, strictness was to be defined in singling out all relevant causes or events which could prevent a nomological effect from obtaining (see Davidson 1980: 219). In this sense, probabilistic laws can be strict: A probabilistic law's consequence may still fail to be brought about by the obtaining of all antecedent conditions (this is what "probabilistic" means, after all), but the law remains strict if there is no additional antecedent which could be cited to account for the non-manifestation of the consequence. That is, if a probabilistic law has the form "F → G", specifying that if F

[35] Neural breakdowns, which are the subject of biology, chemistry and physics can currently (i.e. without appropriate bridge laws (cf. Nagel 1961: ch. 11, Sklar 1967: 118-121) between neural and mental states) be cited as psychological explanations only insofar as they are relevant to the obtaining of a psychological effect, without themselves relying on psychological laws. If specific neurological facts are found to reliably correlate with intentional capabilities (such as forming beliefs, desires, intentions), they become part of the evidential basis of ascription (see section I.7.1). For example, if specific lesions have been found to correlate with an impairment in forming intentions, someone's suffering from this type of lesion plausibly counts as evidence against their having an intention. Whether or not neuroscientific facts outweigh other evidence obviously depends on the totality of the available evidence: for instance, it is hardly justified to judge someone as not depressed when they are showing all the behavioural signs of depression, but lack the neural ones. However, this potential conflict only mirrors the sort of conflict that can *always* arise between different forms of evidence for mental ascriptions.

[36] Whether anything that could happen to water at 0° C, such as my stirring it, could be subject to physical law is at least an open question. That is, even assuming the completeness of physics, my stirring water should turn out to supervene on physics, but it is not exactly my stirring that explains the water's not freezing, since stirring is not a physical kind. And, while often assumed, the completeness of physics and its relation to non-physical laws is itself hotly debated (cf. Papineau 1991, Gillett & Loewer 2001, Gillett [unpublished], Yates 2009, Lowe 2000, Wachter 2006, Mendonça 2010, Stapp 2009, Tiehen 2015, Vasilyev 2009, Montero 2006, Larmer 1986; also see Morrison 2000).

obtains then G will obtain with a given probability, then the law is strict if nothing but the absence of F could explain why G does not obtain.

According to one form of physicalism, a physical law need not exlude all non-physical hindrances, because physics is causally closed (see footnote 36). This notion of strictness is not true for mental laws, since there can be non-mental causes interferring with mental effects. However, we have good reason to believe that if all non-mental hindrances have been excluded, and all relevant mental causes have been cited in the mental law's antecedent, then its consequence would at least probabilistically come about, and the probabilistic consequence is in fact *best explained* by invoking the relevant mental cause(s). In this sense, mental laws are relatively strict – which means there are some phenomena which they explain best, even though, say, ultimate physics may also ideally explain them. It also means that there is the possibility of singling out all relevant mental causes or events which could prevent a nomological effect from obtaining, but not all possible non-mental causes. Relative strictness is in fact true of all good sciences except ideal physics. For example, an astronomical cause (an asteroid impacting earth) may have a biological effect (on life on earth), without there being any possibility of restating the asteroid's impact in biological terms, or the damage to life on earth in astronomical terms.

I.7. Intentional Ascriptions

I.7.1. Evidence for Mental Ascriptions

For the sake of simplicity, I will restrict my claims in this subsection to explicitly apply to propositional attitudes only, that is, to ascriptions of intentional mental states which have the form "X Ms that P", where X stands for a person, M for a verb expressing a psychological attitude (such as a belief, desire or intention), and P for a proposition (see section I.1). What I say may also apply implicitly to non-propositional cases, and these include cases which can be analysed analogously, but cannot be stated in terms of that-clauses for grammatical reasons, such as emotions ("I hate heavy traffic") or perceptions, but I will not explicitly argue that they do. Emotions and perceptions clearly have intentional objects, and these can often be easily turned into the objects of propositional attitudes ("I hate that traffic is heavy"; see section I.1). But even if they cannot, the underlying logic should not be too different. However, depending on your preferred inventory of mental states, you might

also want to admit non-propositional states which differ substantially from those I invoke. Presently, I will not consider those, and my claims might not at all apply to them.

As has been noted Cummins, the assumptions underlying intentional explanations "are seldom if ever made explicit, just as one does not make explicit the mechanical assumptions about springs, levers and gears that ground structural explanations of a mechanical machine. Everyone knows that beliefs are available as premises in inference, that desires specify goals, and that intentions are adopted plans for achieving goals, so it doesn't have to [be] said explicitly (except by philosophers)" (Cummins 2000: 127). And as Lewis points out, "[t]he theory that implicitly defines [belief, desire, and meaning] (...) must amount to nothing more than a mass of platitudes of common sense, though these may be reorganized in perspicuous and unfamiliar ways. Esoteric scientific findings that go beyond common sense must be kept out, on pain of changing the subject" (Lewis 1983b: 112). Consequently, you will not find a technical manual for mental ascriptions here (which Cummins says we all follow implicitly), but rather an investigation of what sort of meta-psychological theories our common practice of ascription commits us to. Mainly, this chapter is supposed to highlight those aspects of ascriptive practice which serve to establish, reinforce and/or clarify the connections between psychological states and intersubjectivity, behaviour and symbolic representation and matters of meaning in general.

But for now, here goes the rough sketch of what a guide to ascriptive practice would look like: The most basic evidence for the ascription of mental states are (a) behavioural cues connected to general attitudes and (b) evidence pertaining to the specific directedness of these attitudes. For example, there are universal behavioural signs for desiring that something be the case (such as looking forward to it, being fixated on it, being uneasy unless it transpires, being relaxed or happy when it does, etc.). Emotionally-laden attitudes are the easiest to spot behaviourally (and as remarked, they may be endowed with propositional or nonpropositional content – I can loathe *that* it rains or I can simply loathe *the rain*). Yet, behavioural evidence may justify a broad range of ascriptions and need to be subplanted by further contextual evidence (see section II.8.4.5).

As Cummins pointed out, beliefs are available as premises in inferences, so we can infer what someone believes by figuring out what someone must believe in order to do what they do and say what they say (etc.). For example, the mere fact that someone is riding the train already justifies the hypotheses that she believes she is riding the train, that she desires to arrive at one of its upcoming stops, and that she intends to get off at the very same. While some emotions can be assigned even without having a hunch about their intentional object –

we can readily see that someone is happy or angry or sad without knowing what they are happy or angry or sad about –, it makes little sense to do the same with beliefs. It's uninteresting to point out that someone simply "believes" without specifying the object of said belief, while pointing out that they're happy or angry or sad is quite informative. This, again, is due to emotions being closely tied to typical behavioural cues, whereas belief is tied to what is being held true. To say that someone is angry makes sense against the backdrop of their possibly not being angry all the time, while there is no state of "not holding true anything" that would be of widespread explanatory use in social interaction. So, it is the intentional object of a belief that is primary to ascribing a belief, while no such thing is necessarily true of emotions and similar mental states that come with behavioural stereotypes and rather overt bodily and somatic states. As in the case of riding a train, what is held true by someone is fundamentally ascribed by what we think they are likely to know given their environment, given their perception of the environment, and given the decisions they must have made and the intentions they must have had given the actions we observe them carry out.

Cummins' pointing out that desire-contents "are available as goals, i.e., conditions whose satisfaction ends processing cycles" (Cummins 1991: 14) hints at potential evidential bases for ascribing desires and their contents: there is observable evidence for pursuing a goal, being invested in pursuing it, or being emotional about its pursuit; and the content of the respective desire can be reconstructed from the pursuit (even though goals are not necessarily overtly pursued). This is not to imply that psychological attitudes and their contents can be directly reduced to (dispositions to) overt behaviour, or that they can be assigned one by one; but there are comparably overt states which are more likely candidates for being assigned more straightforward attitudes and contents, and these can serve as tentpoles for a comprehensive theory of a person's mental states. (These considerations concerning holistic ascriptions will be further pursued in I.7.4.)

Apart from these direct sources of evidence, which depend on observation of someone's behaviour and the relation to their environment, we often rely on many indirect ones, such as:

(1) self-reports about mental states, as well as their derivative forms, such as relayed/second-hand self-reports,
(2) second- or third-person-reports of direct evidence,
(3) inferences to what mental properties are usually connected to (i.e. inferences supported by psychological laws),

(4) common social determinants of mental states, inferences to common mental causes or consequences of social properties (such as: this person was brought up at a Catholic school, so they probably know about the doctrine of the Holy Trinity),

(5) inferences to common diachronous developments, i.e. mental states which are lawfully or rationally implied as a subsequent consequence of another mental state (such as the natural progression of anger or sadness, the progress of stances toward new acquaintances or relationships etc.),

(6) conditions of rationality: If we have evidence for mental state A, and mental state B is rational if mental state A is held, then we are justified in concluding that mental state B is, ceteris paribus, likely to be held (the degree of likelihood depending on additional behavioural or environmental cues, general judgments of the subject's rationality, of how catastrophic the lack of B would be for their status as a reasonable person, etc.).

I.7.2. Do Mental Ascriptions Refer to Private States?

Psychology is, by its very name, the science of the soul. In Western culture, the concept of the soul stands in a long tradition of being associated with metaphysical entities and allusions of divinity – with what today's psychologists would not accept as scientific at all. Early Greek notions of the soul construed it as the entity whose possession makes the difference between being animate and inanimate, between being alive and being a lifeless object. Consequently, we can find ideas such as the immortality of the soul discussed by Plato (cf. his *Phaedo*, 70b, 76c, 78b-80b). This aspect has been adopted by Christianity, whose dogmas are closely intertwined with hundreds of years of Western philosophical tradition and eventually enlightenment, from which psychology eventually emerged in the late 19th century. In the Christian tradition, the notion of the soul is closely intermingled with notions of the divine, of the soul as being a metaphysical entity: it comes from, goes to, or exists in a realm that is beyond the physical. Consequently, some pressing questions about the soul's connection to the body have arisen,[37] which cast their long shadow even over today's debates,

[37] Cf. St. Augustine: *On the Trinity*, book 6, ch. 6 and St. Aquinas: *Summa theologica*, part 1, question 76, art. 8; *Quaestiones disputatae de anima*, art. 10; *Summa contra gentiles*, book 2, ch. 72.

such as those concerned with the causal powers of mental properties (see e.g. Jackson 1982, Kim 1993).

Famously, discussing such conundrums also makes up a substantial part of René Descartes' philosophical body of work, and understanding some of his considerations is particularly instructive for one of our present problems. The 17th century philosopher is widely seen as the first "modern" philosopher, and his modernity is evident in his scientific approach to mathematics, nature and the human body: According to Descartes, a purely mechanical account of human physiology, devoid of the notion of the soul, can explain much more than what the scholastic philosophers had thought possible, namely

> "the digestion of food, the beating of the heart and arteries, the nourishment and growth of the limbs, respiration, waking and sleeping, the reception by the external sense organs of light, sounds, smells, tastes, heat and other such qualities, the imprinting of the ideas of these qualities in the organ of the 'common' sense and the imagination, the retention or stamping of these ideas in the memory, the internal movements of the appetites and passions, and finally the external movements of all the limbs" (AT XI: 201, CSM I: 108).

Yet, Descartes expounded a dualistic conception according to which the soul was made of a substance radically different and separate from the physical. Unlike Ancient Greeks and scholastics, who would hold the soul responsible for the animate aspect of the body, he would trace back mainly consciousness, subjective phenomenological experience and intellectual powers to the workings of the soul (cf. Bennett & Hacker 2003: 26), and he would locate its interaction with the body within the brain's pineal gland:

> "The part of the body in which the soul directly exercises its functions is (…) the innermost part of the brain, which is a certain very small gland situated in the middle of the brain's substance and suspended above the passage through which the spirits in the brain's anterior cavities communicate with those in its posterior cavities. The slightest movements on the part of this gland may alter very greatly the course of these spirits, and conversely any change, however slight, taking place in the course of the spirits may do much to change the movements of the gland" (AT XI: 351, CSM I: 340). "My view is that this gland is the principal seat of the soul, and the place in which all our thoughts are formed. The reason I believe this is that I cannot find any part of the brain, except this, which is not double. Since we see only one thing with two eyes, and hear only one voice with two ears, and in short have never more than one thought at a time, it must necessarily be the case that the impressions which enter by the two eyes or by the two ears, and so on, unite with each other in some part

> of the body before being considered by the soul. Now it is impossible to find any such place in the whole head except this gland; moreover it is situated in the most suitable possible place for this purpose, in the middle of all the concavities; and it is supported and surrounded by the little branches of the carotid arteries which bring the spirits into the brain" (AT III: 19–20, CSMK 143). "Since it is the only solid part in the whole brain which is single, it must necessarily be the seat of the common sense, i.e., of thought, and consequently of the soul; for one cannot be separated from the other. The only alternative is to say that the soul is not joined immediately to any solid part of the body, but only to the animal spirits which are in its concavities, and which enter it and leave it continually like the water of river. That would certainly be thought too absurd" (AT III: 264, CSMK 162).

Apart from some memories which he thought of as being partially stored in the pineal gland and in the muscles (AT III: 20, CSMK 143; AT III: 48, CSMK 146), he also conceived of another kind of memory which is "entirely intellectual, which depends on the soul alone" (AT III: 48, CSMK 146). Descartes' criterion for determining whether a function belongs to the body or soul was this:

> "anything we experience as being in us, and which we see can also exist in wholly inanimate bodies, must be attributed only to our body. On the other hand, anything in us which we cannot conceive in any way as capable of belonging to a body must be attributed to our soul. Thus, because we have no conception of the body as thinking in any way at all, we have reason to believe that every kind of thought present in us belongs to the soul. And since we do not doubt that there are inanimate bodies which can move in as many different ways as our bodies, if not more, and which have as much heat or more […], we must believe that all the heat and all the movements present in us, in so far as they do not depend on thought, belong solely to the body" (AT XI: 329, CSM I: 329).

For Descartes and many philosophers since, the notorious legacy of this dualistic metaphysical view consisted in a nagging skepticism born out of the so-called "problem of other minds". Solipsism, the view that nothing exists beyond one's own consciousness, is intertwined with not finding a solution to the other-minds-problem. The latter essentially consists in the mystery how anyone can know about the contents (or even existence) of another's mind, if the only mind she has immediate access to is her own.[38] Here, "access" is construed as introspective access: a direct, immediate form of awareness of mental content

[38] There are several variants of solipsism and the other-minds-problem; here, I am only talking about an epistemic variant, which, strictly speaking, means that there is no way to verify whether something beyond one's own mind exists, but does not need to entail an all-out denial of the existence of an external world.

which is not inferred from further evidence; the form of access someone has to the content of their own consciousness. While post-Freudian psychology, as well as recent studies on heuristics, biases and ensuing confabulation (cf. Sie & Wouters 2010: 126-128)[39], have familiarised us with the concept of mental content which we are systematically unaware of, it is still true that *if* we are aware of some content of our minds, then we are directly aware of it without consulting further evidence since "we seem intimately acquainted with our own minds" (Heil 2000: 131). "The person who has a desire (or want or belief) does not normally need criteria at all—he generally knows, even in the absence of any clues available to others, what he wants, desires, and believes" (Davidson 1980: 15). And so, we are aware of our own mental states in a way that is different from our awareness of the content of others' minds: there is an asymmetry between knowing one's own pain and knowing that of others, knowing one's own thoughts and knowing those of others, knowing one's own desires and knowing those of others, and so on. Because if we can ever know about others' mental states, then we know about them indirectly: by interpreting observable evidence.

While the fact that this asymmetry exists in some form can hardly be disputed, the Cartesian view goes wrong in supposing that mental states are essentially radically private states, and that the only epistemically justifiable access to them is the first-person access, it being the only *direct* or *immediate* form of access. On such an account, accepting that any access to someone else's mental state can only happen indirectly results in nagging doubts: it may be possible that every single time I ascribe mental properties to someone else I am in fact wrong, and thus it ultimately appears conceivable that no one beside me has ever actually had a mental state and that the only mind that exists is my own. That such solipsistic considerations are entailed by this theory is perhaps its most fundamental weakness.[40]

While the Cartesian account captures some important facts about mental states, it also conveniently neglects others, and draws many unwarranted conclusions (for a broader criticism see Ryle 1949: ch. 1). It is right about there being an asymmetry between first person and second/third person mental state ascriptions; it is right about the fact that when we ascribe mental states to ourselves we often do so immediately, without consulting evidence; and it is right about the possibility of our third-person ascriptions being wrong in each single instant. What it conveniently neglects to acknowledge is that our first-person ascriptions are far from infallible as well, and that they can also work much like our third-person ascriptions do: sometimes, we do not actually know what we really think, plan or desire until someone

[39] Also see Tversky & Kahnemann 1974, Gigerenzer 2008 and Sunstein 2005.
[40] For a modern version of Cartesian solipsism see Putnam's "brain in a vat" scenario (Putnam 1981: 1-21), which was popularised by the 1999 movie *The Matrix*.

points it out to us, or until we come across compelling evidence. It neglects to mention that our first person ascriptions come with the same built-in capacity for being wrong as third-person ascriptions do: that is, each first-person ascription which we can think of as a plausible hypothesis rather than an immediate expression can also be wrong. This may not apply to expressions of inner states such as "I have a toothache", which, if expressed sincerely, directly expresses an immediate feeling of pain; but it does apply to many common intentional ascriptions: we can ascribe any number of beliefs, intentions and desires to ourselves and be wrong about them. We can be wrong about the fact that we believed, intended or desired what we thought we did, we can be wrong about the content of our mental states, and we can even be mistaken in our use of the concepts with which we express them.

The easiest cases of mistaken first-person ascriptions are certainly those that are concerned with a wrong use of concepts, or self-ascriptions of knowledge: A few years ago, I may have believed that supervenience was an implausible stance to take on the mind-body relationship; but now I know that my belief was not in fact a belief about supervenience, because I actually did not properly understand what supervenience was (and I only thought I did). So, we can have beliefs which we later find out to be mistaken because we did not get their content right; and we can believe that we know something to be the case and later find out that we did not know at all.

Trickier ways in which we can be mistaken are concerned with beliefs we ascribe to ourselves based on circumstantial evidence, such as someone's taking the hypothesis "I have read several of Murakami's books, so it may well be the case that I like reading Murakami" as a good reason for believing that she likes reading Murakami. If this person were to later reconstruct that the real reason for her repeatedly reading Murakami was that his books were most convenient to come by, she might revise her belief about liking to read Murakami. Such cases are based on the role of mental states as action explanations and not so much on any immediate feeling or awareness we might have that introspectively connects us to their content. Often, we simply know better what we do or have done than what we believe or desire, and so the latter may be inferred from the former.

These considerations suggest that while there is an asymmetry between first- and third-person ascriptions, there are many cases in which they work analogously or similarly, and that we had better not take those cases which are dissimilar as exclusively delineating the characteristics for what we adopt as our view of mental states. My stating that, say, Tom is mad can only ever be justified if Tom's madness is an intersubjectively evaluable fact. If the only facts about mental states are of a subjective nature, and if the only access to them is a

direct one through the inspection of one's own mind, of an immediate acquaintance with one's mental states by way of introspection, then this view quickly leads us to said skeptical problem: How can I really know that anyone beside myself has mental states? If we are Cartesians, the only possible answers seems to be: We can't and we don't. This skepticism about other minds is thus intimately intertwined with the view that mental states are exclusively inner, private, immediate states.

What the Cartesian picture is resistant to legitimise is the empirical method of gathering information about someone's mental state. I cannot know whether Tom is mad right now without knowing some facts about the world which I can only find out by way of empirical observation. It may not even be required of me to *directly* observe Tom; maybe I can also find out about Tom's mental state by inquiry (which is just to say: by observing someone or something else). What matters is that I have no direct access to Tom's mental state, no access that could be completely independent of empirical data. We cannot know anything about someone else's mental state without observation (where "observation" is used in the rather broad sense I just established, namely as: gathering data about something beyond the immediate content of my own mind). Of course, in order to gain this data, I will have to make it accessible to my mind, thus eventually making it potentially *also* immediately available to me through introspection. But this point may lead us onto to the wrong route, namely to supposing that subjective knowledge (i.e. immediate, introspective knowledge about one's own mind) is the basis for all other knowledge: for intersubjective knowledge (the knowledge about other minds) and for objective knowledge (the knowledge about empirical facts in the world; compare I.7.4).

How do Cartesians, who would suppose that subjective knowledge is the form of knowledge based on which they must establish other forms, go about in trying to justify the ascription of other people's mental states? Ad hoc, plausible strategies could involve a justification from similarity – *I observe objects which look and behave similarly as I do, so perhaps their properties are similar to mine; Having mental states is a property of mine; hence, I can assume that those looking and behaving similarly also have mental states* – or justification from causal relations, "arguing from one effect back to its cause and out again to another effect" (Jackson 1982: 134) – *my mind seems to me to be the cause of my actions; these actions have observable properties; I can observe these properties in events not caused by my mind; perhaps these are effects caused by other minds* –, but it is not at all clear how the relevant inferences are themselves justified (this unclarity is marked by my use of "perhaps" in either inference). Given the Cartesian picture, what norm governing mental

ascriptions can itself justify that the ascription of mental states should hinge on non-mental properties? Even if there is one such norm, the ascriptions of mental states to others always seem considerably weaker than ascriptions based on immediate acquaintance. In fact, in the case of immediate acquaintance, we are tempted to not speak of an "ascription" at all, but rather of "just having" these mental states: of directly knowing that we have them and what they consist in. So, ascriptions are based on evidence, but in the first-person case, there is no evidence necessary at all: we do not infer that we are angry, we *just know it*.

These considerations are prominently on display in Descartes famous "cogito ergo sum" (cf. Descartes 1965): The immediate acquaintance of my mental states – in his case: the fact that *I am thinking* (*cogito*) – comes with absolute epistemic certainty: it is self-evident. Thus, the logically weaker claim that *I exist* (*sum*) is certain as well. But the downside to Descartes' being able to cash out said view in exactly this way is the idea that subjective access to mental states is the benchmark for what we can know about them, and it instantly devalues all other forms of access, and all other forms of knowledge. So, if we want to acknowledge that the Cartesian *ego* is not potentially the only being with a mind in the entire universe we'd better come up with an alternative view.

Before we get to that, let me briefly preempt a potential misunderstanding: the claim that mental states are "inner states" in the sense discussed here should not be misunderstood as directly pertaining to any claims about how anything *within the human body* relates to mental states. "Inner" is only used to distinguish between what belongs to the mind and to what is external, and thus between immediate access and empirical access. The claim that mental states are literally "internal", i.e. states which are characterised by what is *within the human body*, is distinct. So, the problem of not having direct epistemic access to our peers' mental states should not be confused with any problems about not having epistemic access to their inner physiological states (after all, such troubles could be remedied by invoking x-rays, fMRI, surgery or the like). Neither does the view that we have a form of immediate acquaintance with our own mind amount to a claim about having direct access to our own physiological states, and while somatic states (such as those marking excitement, anxiety, exhaustion and the like) do stand in a causal and/or informative relation to the state of our mind (cf. Damasio 1996), the Cartesian sort of introspective access we have to our mind does not amount to a rundown of our physiological facts or somatic states. Conversely, adopting or rejecting this view will not directly impact one's view on physiological facts, including the physiological foundations of our mental states. For example, I believe the "inner states" view

to be inadequate while holding on to the view that our mental states are partly based in our physiology.

I.7.3. Are Intentional Terms Behavioural?

The classic view I just laid out has treated mental properties as radically internal, as something that was only immediately accessible through the intellect's turning on itself in introspection, and as removed from observable evidence. In the mid-20th century, behaviourism caused the pendulum to swing into the other direction. "The behaviorists took [developing a scientific psychology that was on a par with the physical sciences] (…) to require the rejection of nineteenth-century introspective psychology, its method of introspection and "its subject matter *consciousness*"" (Sullivan 2014: 54 f.). Behaviourism so conceived urged that we should solely concentrate on what's observable, methodologically allowing *nothing* but behavioural evidence to justify psychological attributions: "consider only those facts which can be objectively observed in the behavior of one person in its relation to its prior environmental history. If all linkages are lawful, nothing is lost by neglecting a supposed nonphysical [i.e. mental] link" (Skinner 1974: 14). According to behaviourism, some "mentalistic things or events (…) can be 'translated into behavior', others discarded as unnecessary or meaningless" (ibid.: 19).

Just like Cartesianism, behaviourism got many things about mental states right, while conveniently neglecting others and drawing some unwarranted conclusions. The behaviourist is certainly correct in stressing the importance of observability and empirical methods for mental ascriptions (which Cartesianism treated so disrespectfully), and in holding that what is observable in the context of mental properties will in common situations amount to behaviour. As we have learned from Davidson's, Fodors and Sellar's quotes, mental states are often taken to be abstract and unobservable (see I.5), so observing behaviour should be the next best thing.

From what has been said about behaviourism so far, it seems to imply that a type of human behaviour can be explained without recurring to mental properties, namely "those facts which can be objectively observed in the behavior of one person in its [lawful] relation to its prior environmental history". If that were true, then for the postulation of mental states to explain anything at all, we should only (have to) introduce them in those situations in which they go beyond behaviourism, i.e. in which the details of a person's "prior environmental history" do not by themselves lawfully link to behaviour. So, introducing mental states would

make sense if prior environmental history fails to fully explain objectively observed behaviour. Any case in which someone's behaviour cannot be fully derived from her prior environmental history would thus qualify as a potential object of mental explanation.

However, while some interaction of mental states which fails to be fully stated in terms of prior environmental history is at times needed to properly account for an explanation of some instances of behaviour (and I will get to such cases shortly), the underlying picture gets something essential wrong. Because mental states do not merely serve to explain what environmental histories cannot, but rather, as I have shown in section I.6, they explain how environmental histories and behavioural reactions are systematically linked in the first place. That is, a behaviourist theory of higher human cognition might superficially get rid of mental states by admitting only such laws as "whenever their sibling dies, sane persons react with sorrow", where the death of the sibling and the person's sanity, for the sake of the argument, is stated in terms of the person's prior environmental history, and the sorrowful reaction is stated as behaviour. But should we say that such a theory has really gotten rid of mental states in favour of behaviour? I am inclined to argue that the fact that under said circumstances someone reacts with sorrow is to say something essential about her mental properties: It says something about what she believes (namely, that her sibling died), about what she desires (namely, the well-being of her sibling), and how this belief and this desire interact to cause a reaction. And even if as much as I have conceded can be stated in non-mental terms, it is the connection between the world, her perception of it, her belief that is based on the perception, the interaction between the resulting belief and her desire, and the consequent feeling without which we could not explain the connection between prior environmental history and behavioural effect. Because the following psychological law remains true: Were it not for all these mental properties, then the behavioural effect would not occur (or, if it did occur, it would be unexplained).

But what except a mental apparatus could said lawful linkage be based on? Even if there were an extremely robust statistical relation between a persons' sorrow and the prior death of their sibling, it would by itself not amount to a law, since we usually (but certainly in this kind of scientific context) speak of laws when the connection between two things (or events) is causal (see I.6.2). In fact, since we are prone to taking statistical connections as indicators for some (direct or indirect) causal connection, we would be bewildered if there were absolutely no causal connection that could explain why two events show a robust and rather exclusive statistical connection (i.e. where a particular kind of event reliably and exclusively follows another kind of event). And that we would be bewildered just means that

we would not take the statistical connection itself to already be of explanatory value. Rather, its value is heuristic, tracking potential causal relations. In the case we are presently discussing, it is only because we can in fact *understand* how someone comes to feel sorrow upon the death of her sibling that we can postulate the lawful linkage between prior environmental history and behavioural effect. That we can understand this means that there are facts explaining the linkage. And what explains it is that we know that there is not just a statistical relation between the death of someone's sibling and her sorrow, but that there are instantiations of psychological laws in which one is (ceteris paribus) a cause for the other.

Now, the behaviourist could reply that in her theory, some robust statistical linkages simply reflect "behaviourist laws", which may be all sorts of things except mental – or that the relevant psychological laws *are* in fact just behaviourist laws. The difference between these laws and mental laws would be that in behaviourist laws, the external cause (the sibling's death) and the behavioural effect (sorrow) would themselves be lawfully linked, whereas mental laws rather hold that the *belief* about the external event is what causes an effect which is mental *and* behavioural – sorrow being a state of mind that comes with a behavioural signature. Still, this move on the behaviourist's part would not suffice, as such behaviourist laws still lack explanatory value, since there is nothing in the behaviourist theory which would explain in virtue of what they hold (and how could there, when clearly what stands between an external event and a behavioural reaction is not itself behavioural, but cognitive?). The question "why does she show sadness-behaviour as a consequence of the death of her sibling?" cannot merely be answered by "because the two events are lawfully linked" or "because sadness behaviour usually occurs after the death of a sibling". In mental theories, on the other hand, we can say that under favourable conditions, the external event – the death of a sibling – causes the respective belief, and the belief, interacting with the desire for the sibling to be well, causes sadness, which in turn has behavioural expressions or consequences. So the behaviourist theory lacks an explanation the mental theory doesn't.

But even beyond this explanatory weakness, note that the behaviourist cannot even admit notions such as "being informed about (their sibling's death)" into her theoretical vocabulary. Clearly, prior environmental histories do not by themselves explain the behavioural reactions the behaviourist aims to explain: someone's sibling can very well die in the vicinity without the former feeling sorrow at all, merely by not being informed about it. Even a rat in the "Skinner Box" has to be somehow informed about its reward in order for conditioning to work; no behavioural reaction to an environmental event can occur without the subject's being at least in the most minimal sense informed about it. The behaviourist can

try to amend her laws by replying that the rat's gaze must be directed to the reward at some point, but of course that would beg the question why the gaze is a relevant factor, if not for mental reasons (the same applies to behaviour connected to attention and perception, and many complex forms of learning – explanatory gaps which eventually brought about behaviourism's downfall and ushered in the cognitive turn; cf. Sullivan 2014: 55). Mental laws can explain this fact very well: again, it is the belief that causes the sorrow, and such beliefs depend on being informed about external circumstances. Behaviourist amendments about gaze direction and the like will only serve to point out a gap which mental explanations were made to fill.

Thus, if we either want to fill this gap, or lend said behaviourist laws any explanatory power, or both, we must already assume a mental apparatus. To be sure, mental laws can predict behaviour, and behaviour may serve as the prime evidence for mental ascriptions. But while they can in some instances be formulated in purely behavioural terms (i.e. in those not covered by the example above, such as more complex forms of learning), *even then* all of their "lawful linkage" is due to mental properties. As noted, what stands between an external event and a behavioural reaction is what goes in someone's head, and, among other things, it's these goings-on which mental state ascriptions aim to capture. Thus, it is not at all necessary that we come up with cases in which someone's behaviour *cannot* be fully derived from her "prior environmental history" for introducing mental states into our theory (although there *additionally* are such cases). We find psychological laws at work even when we can infer someone's behaviour from nothing but prior environmental histories.

In any case, behaviourism took a commendable stand against an unwarranted focus on "inner" determinants of behaviour which unduly neglected external determinants and against using introspection as a psychological method to access the former: "By directing attention to genetic and environmental antecedents, [methodological behaviourism] offset an unwarranted concentration on an inner life. It freed us to study the behavior of lowers species, where introspection (…) was not feasible" (Skinner 1974: 16). If we had stuck to the Cartesian picture, elevating mental properties to something radically internal, something only the "self" can have certain access to, then we would never have gained an adequate view of mental properties.[41]

Note that we can find matters of psychology to be intertwined with matters of semantics: On the classic Cartesian model, where the soul was seen as an entity beyond the physical realm and matters of inner experience and intellect could be relegated to this realm,

[41] For a reconciliation of a weaker form of behaviourism with the kind of philosophy of mind which I am partial to see Sellars 1997: 98-107 (§53-59).

the corresponding semantic theory *must* have held that the meaning of terms referring to these mental states must comes from beyond too. So, mental state ascriptions could only be made because they referred to ideas in the mind, and these ideas did not necessarily have to be learned, but came from "beyond" as well (and there is a certain analogy to views which postulate internal mental objects as referents of intentional states, see I.3). But once we change our views about semantics, abandoning the notion of such Platonic ideas – ideas which are not learned, but remembered –, the mental theory must change accordingly. If the most plausible construal of concepts is as things that are learned through social interaction, which are subject to rules of public discourse, and meaning can no longer simply fall from the Platonic Heavens but is acquired by associating stimuli (such as the perception of symbols) with objects, contexts or other concepts, then mastering concepts requires their associated meaning to be publicly accessible. This is the kind of construal I am going to elaborate on in the following section.

If mental state ascriptions are intertwined with public criteria, then there is no good reason to doubt that they are readily intersubjectively available. Consequently, any criteria for ascribing mental states beyond immediate self-ascriptions (such as "I have a tooth-ache") must be based on observable evidence. That is, if the mental state ascription "Tom is angry right now" is fit to be viewed as referring to a fact, then the evidence for justifiably establishing this fact must be empirical. What empirical facts about mental states do we have access to? As pointed out in I.7.1, these will often be *behavioural* facts: "We know about other minds by knowing about other behaviour, at least in part. The nature of the inference is a matter of some controversy, but it is not a matter of controversy that it proceeds from behaviour. That is why we think that stones do not feel and dogs do feel" (Jackson 1982: 134).

When Jackson says that behaviour only partly informs our knowledge, I take him to allow that there may be, say, ways of transmitting information about mental states which is not behavioural itself. For example, someone may utter a justified mental state ascription, and someone else might record it. Given that I have good reason to believe I am witnessing the recording of a justified mental state ascription, I gain knowledge about the fact expressed by this ascription, even though the way I gained it is witnessing a recording. The recording might consist in a recorder's transducing acoustic signals. The recorder's transducing acoustic signals is not behavioural. Thus, I did not gain knowledge about the mental state ascription by way of behavioural facts. However, the recording can only be reasonable grounds for my gaining knowledge if it can be traced back to behavioural facts. In fact, there can be no

reasonable grounds for gaining knowledge about someone's mental state(s) without the grounds being at least indirectly behavioural. There may be all sorts of transmitters and reports inbetween the original behaviour and my gaining knowledge about the subject's mental state, but they could not report any non-behavioural facts. That is, either they are reports about mental states; then they are informed by behaviour. Or they are reports about behaviour; then they can be grounds for mental state ascriptions. Or they are inferred from contextual information, which depends on a notion of what kind of behaviour would be appropriate given the respective context.

Sometimes, matters are less straightforward, as in the case of a self-report: If Tom tells me that he is angry, then his telling me that he is angry is both a mental state ascription as well as behaviour. Sometimes, a liar's report that he was not lying when he stated that X can be grounds for calling him a liar (namely when we know that X is not the case). Similarly, someone's denial of a mental state ascription can, given other evidence, be grounds for ascribing this mental state to him. If Tom is gritting his teeth, grimacing demonically and stomping his feet while telling me that HE IS NOT ANGRY AT ALL, then I will not only disregard or overrule his self-report in face of the rest of the evidence; rather, since I know that Tom tends to be in denial when angry, I will explain his self-report as a behavioural consequence of his mental state, namely anger, and thus as additional grounds for ascribing anger to him.

This firm connection between mental state ascriptions and observations of behaviour is not merely rooted in what seems like a regrettably restrictive fact, namely that we are often in no position to observe *more* than behaviour. Of course, some will say, we are currently attempting to rectify just that by looking into people's brains! However, the more significant root is that mental states are theoretical terms which are largely used to explain behaviour; that is, even if we can observe them in a way that does not directly rely on observation of behaviour, such as through observations of brain activity, there is no reason to suppose that brain activity will suddenly take the place of any theoretical notions we had derived from systematically different methodical grounds. The value of mental states is not measured by whether they are good predictors for what we can find in our brains, but for the causal determinants of behaviour (compare II.8.4.5). While brain activity will belong to these relevant causes (behaviour is directly caused by brain activity after all), matters external to the brain will also belong to them, because they are part of the characterisation of our abstract states. For example, we can identify witnessing a celebration as the cause of opening a bottle of wine, and we can also trace this action back to its being elicited by specific brain activity.

Yet, the respective brain activity itself will have to stand in a systematic relation to witnessing the celebration, thus making witnessing the celebration a vital part of the causal explanation of the action – even in cases in which we have command of an ideal brain scanning methodology using which we can directly observe and find out everything there is to find out about the relevant brain activity.

I.7.4. The Davidsonian View of Mental State Ascriptions

It is one of the great contributions of 20th century philosophy and psychology to have resulted in a viable solution to Cartesian skepticism and to have shown how mental attributions based on evidence are systematically justified, all the while treating them with the proper respect they were denied by behaviourism.[42] As is often the case, virtue lies in moderation, and it is the moderate position inbetween radical subjectivism, that comes with the threat of solipsism, and radical behaviourism, which endeavoured to take the mind out of psychology, which is most prudent to adopt. On the one hand, we should recognise that attributions of mental states to other persons are systematically justified, in the same way empirical hypotheses are justified: they can be wrong in each instance, but it is nonsensical to doubt the sheer possibility of justifying *any* empirical hypothesis based on observable evidence. For example, there can without a doubt be good evidence for justifying the judgment that a given object is made of stone, but of course this does not imply that every time someone is justified in judging that something is made of stone, she has to be right about it. Analogously, we may judge someone to be angry based on her displayed behaviour, and it is the observability of this behaviour which justifies the hypothesis that she is angry. It may turn out that she actually isn't; but it is nonsense to doubt that a proper cause of behavioural signs of anger is someone's being angry and that, thusly, angriness-behaviour constitutes proper grounds for the justified hypothesis that the person in question is angry. One such moderate stance is Donald Davidson's position. Adopting it, we can also give clearer meaning to the claim that mental states are theoretical entities.

[42] I say "theoretical", because solipsism may pose a pressing theoretical problem, but never a pressing practical problem: No significant amount of solipsists, if any have ever actually existed, have ever consistently behaved as if no external world consisted. As construed in the present context, solipsism is essentially a systematic gap in the justification of (mental) attributions, but which we regularly engage in anyway. (On a side note, Piaget has suggested that infants in fact undergo a solipsistic stage of development and have to consequently "convince" themselves that what they perceive are actually the effects of external objects, cf. Flanagan 1991: 144 ff.. Even if that's the case, said gap in justification is surely an independent matter.)

"Sometimes skepticism seems to rest on a simple fallacy, the fallacy of reasoning from the fact that there is nothing we might not be wrong about to the conclusion that we might be wrong about everything" (Davidson 2001b: 45). That it is true that we could err in any single instant does not establish that we cannot err systematically in all instances at once. The logic of this fallacy is similar to inferring the wrong statement "all Catholics could be elected Pope at the same time" from the true statement "any Catholic could be elected Pope". In discussing the Cartesian view I have already mentioned one root of this fallacy, namely holding that intersubjective knowledge is inferred from subjective knowledge. Davidson argued that subjective, intersubjective and objective knowledge are interdependent: having or being able to have one kind requires being able to have the other two, and none is reducible to any other form. "There are (…) no 'barriers', logical or epistemic, between the three varieties of knowledge. On the other hand, the very way in which each depends on the others shows why none can be eliminated, or reduced to the others" (ibid.: 214).

Davidson bases this view on what he calls "Triangulation" (ibid.: 212 f.). According to this view, the acquisition of language, and consequently, mastering subjective and objective concepts, requires intersubjectivity along the following lines: learning a language means acquiring dispositions to react to similarly perceived stimuli with similar utterances (see Figure 3). Our ability to do so can be explained evolutionarily (ibid.). What is required beyond an underlying cognitive mechanism which produces these consistent perceptions is a social situation in which this association of stimuli and utterances can be learned: "it is only when an observer consciously correlates the responses of another creature with objects and events of the observer's world that there is any basis for saying the creature is responding to those objects or events rather than any other objects or events" (ibid.: 212). Expressions which can be directly bound to non-linguistic stimuli, much as the uttering of "lo, there is a rabbit" can be bound to the presentation of a rabbit-stimulus, are what Quine calls "stimulus meaning" (cf. Quine 1960: 32-36). And linguistic expressions themselves can serve as stimuli which, by way of further conditioning, will elicit more utterances. This form of conditioned meaning will only indirectly be tied to non-linguistic stimuli, and thus any statement that relies on more than just stimulus meaning will be less easy to relate to empirical goings-on. Davidson's analogy is geological triangulation, in which an object is scrutinised from two different points of view; in the geological case, the object's distance or height can be determined (depending on which of the two variables is known), while in language acquisition, the concept's meaning can be determined:

> "Without [a] sharing of reactions to common stimuli, thought and speech would have no particular content – that is, no content at all. It takes two points of view to give a location to the cause of a thought, and thus to define its content. We may think of it as a form of triangulation: each of two people is reacting differentially to sensory stimuli streaming in from a certain direction. Projecting the incoming lines outward, the common cause is at their intersection. If the two people now note each other's reactions (in the case of language, verbal reactions), each can correlate these observed reactions with his or her stimuli from the world. A common cause has been determined. The triangle which gives content to thought and speech is complete. But it takes two to triangulate" (ibid.: 212 f.).

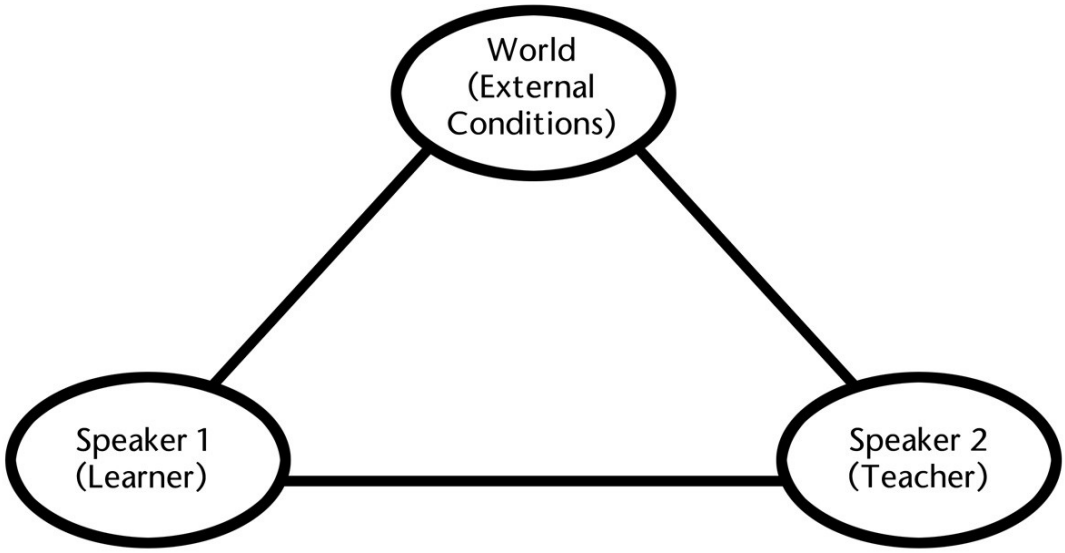

Figure 3: Davidsonian Triangulation.

This triangulation is the basis for ascribing mental content, and therefore symbolic content is public in nature:

> "When we start learning a language, we associate linguistic expressions with our different anticipations and other dispositions on the basis of publicly accessible evidence. (…) [A]s soon as we extrapolate from the perceptual realm into more theoretical domains, the interplay between theory and meaning becomes more pervasive. (…) In these areas it is important that we not regard meaning as something that first exists in our mind and then gets expressed through language. There are no proto-meanings in our mind, as [Jerry] Fodor and many others maintain [see I.5.5]. There are intimate and interesting connections between mind and meaning. But we get a wrong picture of these connections if we fail to take seriously the public nature of language" (Føllesdal 1975: 43).

While meaning is a social matter, it is objective insofar as it is not *determined* socially, or reducible to social facts. Even an ideal complete description of social matters – of how all forms of individual behaviour have been tied to or correlated with environmental cues – underdetermine meaning, much as empirical facts underdetermine theories in the natural sciences (and as Quine and Dennett would argue, the indeterminacy goes even further; see I.6.1). On the Davidsonian view, theories of meaning are supported by empirical evidence, but this relation of support is not one of direct implication. Since theories have a potentially infinite amount if instances they govern, they cannot be reduced to a finite set of evidence. Likewise, no definite meaning can be given for any linguistic entity; for linguistic meaning springs from an ever-ongoing process that consists in associating sentences, or parts of it, with non-verbal stimuli and/or other sentences (cf. Quine 1960: 9 ff.).

It should be noted that accepting a fundamental indeterminacy which goes beyond the way in which theories in the natural sciences are underdetermined by empirical evidence would fundamentally distinguish psychological theories from those in the natural sciences. For instance, dealing with neurobiological methods, just as with methods of physicochemical measurements in general, we would be extremely wary of them if they systematically yielded contradictory results. Spoken in everyday, macroscopic terms: In the place of an apple, there can't also be an orange. It can be next to it, or on top of it, or under it, but not in the exact same place. Similarly, if two measurements of ion-concentrations in the same neuron at the same time would contradict each other, we would assume one measurement to be wrong, and our apparatus to be faulty. Not so with psychological measurements, because all we can ever measure empirically is the evidence for the ascription of psychological states rather than the psychological state itself, and if indeterminacy is true, these states are not singularly determined by it. That is, psychological questionnaires will yield definite statements, or show definite boxes being ticked; our inspection of someone's facial expression may yield a definite grimace, a behavioural description will yield a definite sequence of movements in space and time – but what we make of all of these is subject to a schema of intentional interpretation. I will accept the latter point; but what kind of indeterminacy follows from it I will leave open. If it is merely a Davidsonian one, which is best expressed as the kind of indeterminacy that always exists between using different scales of measurements, then we're on a safer side than if we go with the Quine or Dennett kind (see I.6.1).

In a Wittgensteinian sense, knowing the meaning of a word or sentence is knowing how to use it – knowing the rule according to which it is used (compare Wittgenstein 1953:

§43, also see section I.4.1). And these rules, in turn, cannot be entirely extrapolated from empirically gathered data: At any given time, there is an infinite amount of potential rules which would chime with past linguistic behaviour (compare Kripke 1982 and II.8.2). And the problem is not just one of extrapolating a determinate, independently existing rule to govern linguistic meaning from verbal and non-verbal behaviour, since our peers, who we would like to learn meaning from by extrapolating the rules governing *their* linguistic behaviour, are essentially in the same boat: They, in turn, have learned semantic rules by trying to extrapolate rules from other persons' linguistic behaviour as well. So, in learning language, we extrapolate rules governing others' linguistic behaviour which is in turn guided by extrapolated rules, ad infinitum.

So, knowing the semantics of a language consists in having a theory about the rules that govern linguistic behaviour, and every set of rules that conforms to the evidence constitutes a workable theory. Since the linguistic behaviour that serves as evidence is in no way independent from the semantic theories which are tested in comparison with this evidence, revising a theory can be weighed against revising linguistic behaviour itself (such as by pointing out that a given term has been used mistakenly): "there will always be cases where all possible evidence leaves open a choice between attributing to a speaker a standard meaning and an idiosyncratic pattern of belief, or a deviant meaning and a sober opinion" (Davidson 1980: 257). Which option we choose will also be influenced by considering criteria of theoretical simplicity (cf. Quine 1960: 19 ff.). This whole picture is strikingly akin to theory-formation in the natural sciences: We are confronted with a complex multitude of observable phenomena, and we try to structure, explain and predict them by coming up with a theory (cf. Davidson 2001a: 222). Meaning is the operative aspect in such theories (cf. Davidson 1980: 256 ff.), and in this sense it is fundamentally theory-dependent. There is no "free-floating, linguistically neutral meaning" (Quine 1960: 76).[43]

Now, one might paint the alternative picture that we have some sort of naturally built-in semantic organ which informs us about what meaning certain symbolic entities have. While it is true that some of our intentional attitudes are unconditioned (much as we can viscerally extrapolate the notion of being in danger from the appearance of a roaring lion in our direct vicinity), and that human beings are also equipped with a natural disposition to use symbols in manifold ways, learning certainly does play a crucial role in endowing many entities with meaning. For instance, we are not free to replace the visceral "meaning" a roaring lion has for us with some other stimulus in the way that we are free to introduce the convention of

[43] Here, I urge to take "linguistic" in a broad sense; i.e. as applying to all symbolic systems which carry meaning.

meaning 3.14159... by using the symbol π. A "semantic organ" could not explain such matters of acquired (i.e. conventional) meaning; and since the kind of meaning which is not acquired (i.e. "sparse", see I.4.4) can be cashed out in terms of non-representational causality (see II.4 and II.6), assuming any such organ in order to explain matters of meaning would eventually prove superfluous.

Following Quine, what is important for an individual to acquire a semantic theory is characterised functionally (also compare Cummins 2000: 125 f.). While there has to be a hard-wired basis on which our linguistic abilities can flourish, a potentially unlimited multitude of different (neural) routings could enable us to partake in linguistic practice. That is, having a semantic theory, and being able to speak a language, are functional descriptions which can be implemented in a variety of (neural and other) ways.[44] Primarily because it is not the individual body-state which alone determines meaning, but what intersubjectively follows from it. For instance, "square" refers to any "objective" square, even though in a situation where several observers are standing around a tile, the subjective stimulus is a retinal projection consisting in "a scalene quadrilateral which is geometrically dissimilar to everyone else's" (Quine 1960: 7) retinal projection. Compare also the following:

> "To look deep into the subject's head would be inappropriate even if feasible, for we want to keep clear of his idiosyncratic neural routings or private history of habit formation. We are after his socially inculcated linguistic usage, hence his responses to conditions normally subject to social assessment. [Visual stimulation by way of] ocular irradiation is intersubjectively checked to some degree by society and linguist alike, by making allowances for the speaker's orientation and the relative disposition of objects" (Quine 1960: 31).

Since linguistic behaviour is conditioned to observable stimuli, it is the intersubjectively interesting and characteristic features which take primacy in the characterisation of meaningful expressions, rather than the subjective quality of the stimulus; even though the subjective quality has to be reliably associated with intersubjective features (by being "intersubjectively checked") in order to explain why a given individual can produce

[44] Which is not to say that any neural basis has to be sufficient for knowing a language, but that a potentially necessary neural basis can take any shape that satisfies the specified function. This point is independent of any view pertaining to reductionism: it is compatible with the radical reductionist view that knowing a language entirely consists in having a certain neural structure (namely if there is or can ever be only one neural implementation of the given function) as well as with the moderate view that a neural basis is necessary, but not sufficient (i.e. that non-neural factors also play a necessary role in determining whether a given language is spoken). For the radical non-reductionist, who claims that a neural basis isn't even necessary for linguistic knowledge, this whole issue won't even arise: For her, musings about natural organs and idiosyncratic neural routings are completely beside the point.

the respective meaningful behaviour in the first place. By way of learning and using language, we are conditioned to use intersubjective meanings. We are still able to refer to subjective experience, describing feelings, dreams or pains, but that is only because we and our peers have been conditioned to use the words "feeling", "dream" and "pain" under intersubjectively available conditions (compare I.9.4).

Davidsonian Triangulation is a variant of semantic externalism; but unlike Putnam's (see sections I.4.3 and I.8.3), it does not consist in identifying causal chains between objects and concepts (cf. Putnam 1981: chapter 1), but in developing common dispositions in response to objects that are perceived as similar. Just as in Putnam's case, this guarantees the object's independence of the subject's mental state: thus, mental states can refer to matters of fact which go beyond what goes on in the mind. This establishes the connection between subjectivity and objectivity and the fact that any of our statements can be wrong (cf. Davidson 2001b: chapter 2). In this sense, objective knowledge is not derivable from subjective knowledge, but the two depend on each other.

Taking triangulation as paradigm for the acquisition of language guarantees us the interdependence of subjective, intersubjective and objective knowledge: In order to acquire subjective knowledge – knowledge about the content of one's own mind –, it is necessary to be able to ascribe propositional attitudes in accordance with a psychological theory, which can only be acquired through intersubjective triangulation. In order to know what it means to have a belief, one has to understand that beliefs can also be had by others:

> "Until a base line has been established by communication with someone else, there is no point in saying one's own thoughts or words have a propositional content. If this is so, then it is clear that knowledge of another mind is essential to all thought and all knowledge. (…) Knowledge of the propositional contents of our own minds is not possible without the other forms of knowledge since there is no propositional thought without communication. It is also the case that we are not in a position to attribute thoughts to others unless we know what we think since attributing thoughts to others is a matter of matching the verbal and other behavior of others to our own propositions or meaningful sentences. Knowledge of our own minds and knowledge of the minds of others are thus mutually dependent" (Davidson 2001b: 213).

So, self-ascriptions of mental content are only viable if the concept with which these are ascribed have been learned in the context of triangulated objects, thus requiring an intersubjectively shared external world: "Knowledge of another mind is possible, however, only if one has knowledge of the world, for the triangulation which is essential to thought

requires that those in communication recognise that they occupy positions in a shared world. So knowledge of other minds and knowledge of the world are mutually dependent; neither is possible without the other" (ibid.). Objectivity is also guaranteed in another sense: Mastering the ascription of intentional mental states also requires acknowledging that they can be true or false, depending on states of affairs in an external world. We can call this view about the interdependence of subjective, intersubjective and objective knowledge *Epistemic Holism*, meaning it entails that the ability to acquire any single form of knowledge depends on the possibility of acquiring all forms.

Another basis for establishing that other-minds-skepticism rests on a fallacy is *Mental Holism*, a position which we can also find prominently argued for in Davidson's writings and which has its roots in Quine's views (cf. Quine 1980: 20-46).[45] Mental holism is the view that the ascription of mental properties can and does not proceed one by one, but rather by taking their entirety into account. That is, both truth and meaning of an individual mental state ascription depend on the totality of mental state ascriptions as well as on the available evidence for these:

> "We interpret a single speech act against the background of a theory of the speaker's language. Such a theory tells us (at least) the truth conditions of each of an infinite number of sentences the man might utter, these conditions being relative to the time and circumstances of utterance. In building up such a theory, whether consciously, like an anthropologist or linguist, or unwittingly, like a child learning its first language, we are never in a position directly to learn the meanings of words one by one, and then independently to learn rules for assembling them into meaningful wholes. We start rather with the wholes, and infer (or contrive) an underlying structure. Meaning is the operative aspect of this structure. Since the structure is inferred, from the point of view anyway of what is needed and known for communication, we must view meaning itself as a theoretical construction. Like any construct, it is arbitrary except for the formal and empirical constraints we impose on it. In the case of meaning, the constraints cannot uniquely fix the theory of interpretation. The reason, as Quine has convincingly argued, is that the sentences a speaker holds to be true are determined, in ways we can only partly disentangle, by what the speaker means by his words and what he believes about the world. A better way to put this would be to say: belief and meaning cannot be uniquely reconstructed from speech behaviour" (Davidson 1980: 256 f.).

[45] See Lycan 2000: chapter 8 for an introduction to Quine's semantic theory and Duhem 1906/54 for the historical root of his position.

Davidson derived his main argument for mental holism from the work he had done on decision theory (cf. Davidson, Suppes & Siegel 1957). A central aim in decision theory is to predict and explain decisions made between viable options, and this can be done in terms of a preference for one option over another. Preferences can in turn be explained by two kinds of mental attitudes: The desire to bring about a certain event (in technical terms: the "expected utility", cf. von Neumann & Morgenstern 1944) and the subjective probability which is assigned to the realisation of this event. Let's look at the following example:

> "Suppose an agent is indifferent between getting $5.00, and a gamble that offers him $11.00 if a coin comes up heads, and $0.00 if it comes up tails. We might explain (i.e., 'interpret') his indifference [either] by supposing that money has a diminishing marginal value for him: $5.00 is midway on his subjective value scale between $0.00 and $11.00 (…) [or by the agent's belief that] tails are more likely to come up than heads; if he thought heads and tails equally probable, he would certainly prefer the gamble, which would then be equal to a straight offer of $5.50" (Davidson 2001a: 145).

These two alternative explanations illustrate the fact that preferences are vectors of said mental attitudes: Of "the relative values the chooser places on the outcomes, and the probability he assigns to those outcomes, conditional on his choice" (ibid.). Choices or preferences are the relevant observable evidence we have for the ascription of both subjective probability and relative value (or "utility"), which serve to explain the observable events. "Support for the explanation doesn't come from a new kind of insight into the attitudes and beliefs of the agent, but from more observations of preferences of the very sort to be explained. In brief, to explain (i.e., interpret) a particular choice or preference, we observe other choices or preferences; these will support a theory on the basis of which the original choice or preference can be explained" (ibid.: 146). According to this notion, utility and subjective probability are abstract concepts of a theory which we can use to impose a systematic, holistic structure with inherent explanatory value on observable data, and they are theoretical concepts because they only have meaning in the context of such an explanatory theory.

According to Davidson's view, mental states and meaning make up a vector analogously to that of expected utility and subjective probability: "behavioural or dispositional facts that can be described in ways that do not assume interpretations, but on which a theory of interpretation can be based, will necessarily be a vector of meaning and belief" (ibid.: 148). Furthermore, decision theory and interpretation theory are closely

connected: “it is not reasonable to suppose we can interpret verbal behaviour without fine-grained informations about beliefs and intentions, nor is it reasonable to imagine we can justify the attribution of preferences among complex options unless we can interpret speech behaviour” (ibid.: 147). Likewise, “except in the cases of the most primitive beliefs and desires, establishing the correctness of an attribution of belief or desire involves much the same problems as showing that we have understood the words of another” (Davidson 1980: 237).

However, how are we to tackle the problem that matters of mental states and matters of semantics form one vector which we cannot unravel merely on the basis of observable evidence? Deviously, the meaning we ascribe to someone’s utterance depends on what they believe to be the case, and what they believe depends on what they mean by their utterances. Let’s assume we hear someone say “il pleut”. If we can assume that what she means by her utterance is “it rains”, then this utterance is evidence for ascribing to her the belief that it rains. On the other hand, the fact that it rains can only ever constitute evidence for her meaning “it rains” by the utterance “il pleut” if her belief which she expresses by this utterance is not mistaken. But if our ascribing to her said belief is only justified if we know what she means, and we can only know what she means if we know what she believes – so what are we to do? Again, the only observable evidence we have at our disposal is the utterance “il pleut” and the observation of the utterance’s context, namely whether it rains. But these alone cannot help us untangle what someone means from what she believes. So, how can we enter this circle?

“The interdependence of belief and meaning springs from the interdependence of two aspects of the interpretation of speech behaviour: the attribution of beliefs and the interpretation of sentences” (Davidson 2001a: 195). To start, Davidson proposes that we should take the vector constituted by said observables to represent the attitude of *holding a sentence to be true*: for if we know that someone holds a sentence true, and we also know what the sentence means, then we know what she believes; and alternatively, if we know that someone holds a sentence true, and we additionally know what she believes, we know what she means by it. Crucially, we can know whether someone holds a sentence true without having to invoke semantic knowledge:

> “On the one hand, most uses of language tell us directly, or shed light on the question, whether a speaker holds a sentence to be true. If a speaker’s purpose is to give information, or to make an honest assertion, then normally, the speaker believes he is uttering a sentence true under the

> circumstances. (…) In order to infer from such evidence that a speaker holds a sentence true we need to know much about his desires and beliefs, but we do not have to know what his words mean" (ibid.: 161 f.).

Now we also need to determine one of the two remaining factors: "we must have a theory that simultaneously accounts for attitudes and interprets speech, and which assumes neither" (ibid.: 195). "It makes no sense to suppose that we can first intuit all of a person's intentions and beliefs and then get at what he means by what he says. Rather we refine our theory of each in the light of the other" (Davidson 1980: 258).

Davidson's solution consists in embracing the fact that we cannot help but methodically assume a speaker's utterances to be mostly true. This point goes back to Quine's "radical translation" (cf. Quine 1960: ch. 2): If we have only limited knowledge of a language, we can't help but assume that all sentences which are held true by speakers express true propositions, otherwise we could not even begin to understand them. According to semantic holism, two speakers need to share many beliefs in order to be able to mean the same thing with individual statements:

> "If sentences depend for their meaning on their structure, and we understand the meaning of each item in the structure only as an abstraction from the totality of sentences in which it features, then we can give the meaning of any sentence (or word) only by giving the meaning of every sentence (and word) in the language. Frege said that only in the context of a sentence does a word have meaning; in the same vein he might have added that only in the context of the language does a sentence (and therefore a word) have meaning" (Davidson 2001a: 22).

So, for two persons to mean the same thing when uttering "the sun is high up in the sky", they need to share many beliefs about the sun and the sky. This doesn't exclude the possibility that either could be wrong – but it requires that they can only be wrong in individual cases, but not generally. In order to be able to find out which of their statements are false, we need to find out what they mean – and we cannot do that if we do not assume that, looking at the whole of them, they are mostly correct. That is the essence of holism: assigning the value "true" to as many held-true sentences as possible, and it is a methodological requirement both for ascribing meaning as well as for ascribing intentional mental states. The principle that consists of assuming that someone is mostly right with what they say is called "principle of charity". Its application is necessary to untangle the vector of holding-true into belief and meaning; and its application is also justified, since the possibility of error can only come into

play after determining meaning, and thus assuming general agreement (cf. Davidson 2001a: 200; also compare Lewis 1983b: 113 and Dennett 1987: 19, fn. 1).

The principle of charity unites two sub-principles: the "principle of correspondence" and the "principle of coherence". The principle of correspondence holds that we need to assume general agreement between a speaker and her interpreter:

> "Since knowledge of beliefs comes only with the ability to interpret words, the only possibility at the start is to assume general agreement on beliefs. We get a first approximation to a finished theory by assigning to sentences of a speaker conditions of truth that actually obtain (in our own opinion) just when the speaker holds those sentences true. The guiding policy is to do this as far as possible, subject to considerations of simplicity, hunches about the effects of social conditioning, and of course our common-sense, or scientific, knowledge of explicable error" (Davidson 2001a: 196).

According to this principle, people share most of their beliefs if and only if they are able to understand each other, i.e. if they speak translatable languages (or at least translatable fragments). This ties into the holism of meaning: in order to be able to converse about the same thing, we need to share a whole "web of beliefs" into which the concepts we use are embedded (cf. ibid.: 200; Quine & Ullian 1970).

Apart from maximising correspondence, we also need to maximise coherence: That is, we need to assume that those we interpret are mostly consistent and rational. In order to understand a speaker, "we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying)" (Davidson 1980: 222). This requirement of rationality, the principle of coherence, is as indispensable as that of correspondence:

> "Coherence here includes the idea of rationality both in the sense that the action to be explained must be reasonable in the light of the assigned desires and beliefs, but also in the sense that the assigned desires and beliefs must fit with one another. The methodological assumption of rationality does not make it impossible to attribute irrational thoughts and actions to an agent, but it does impose a burden on such attributions. We weaken the intelligibility of attributions of thoughts of any kind to the extent that we fail to uncover a consistent pattern of beliefs, and, finally, of actions, for it is only against a background of such a pattern that we can identify thoughts." (Davidson 2001a: 159). "[I]f we are intelligibly to attribute attitudes and beliefs, or usefully to describe motions as behaviour, then we are

committed to finding, in the pattern of behaviour, belief and desire, a large degree of rationality and consistency" (Davidson 1980: 237).

It should be noted that the form of rationality that is required here is weaker than our common understanding of rationality (compare section I.6.4): It requires a form of internal consistency, both in terms of mental states (i.e. two openly opposing beliefs should not be held at the same time) and in terms of behaviour (actions should be somewhat consistent with what is believed and desired). There's a pervasive idea that human beings are notoriously irrational creatures, an idea that seems to be supported by much recent psychological research (see footnote 39). However, when accusing someone of irrational behaviour in everyday life, we usually mean that they seem to act against plausible norms governing actions, not that there couldn't possibly be understandable reasons for their actions. (In fact, calling someone irrational is often akin to calling out their reasons as stupid or immoral, not as inconsistent). The Davidsonian version of rationality means the latter: it requires that someone's actions can be explained in terms of *any* reasons which are internally consistent and can be consistently tied to external circumstances (but not necessarily to plausible norms governing actions – they may very well be tied to what we deem immoral, stupid or short-sighted norms).

We may also sometimes lack the cognitive capacity to be consistent even in this weaker sense, but such failures can be amended by simply ascribing an additional mental state that explains the unawareness of the respective inconsistency. That is, it is plausible for me to hold two inconsistent beliefs if I have good reasons for being unaware of how they contradict each other. Someone may also act irrationally by, say, avoiding to board planes because of a fear that it might crash, while regularly playing lotto. Again, amendments can be made to our mental state descriptions which would explain such discrepancies. However, the point remains that our actions and mental attitudes cannot be grossly inconsistent, for then they would be uninterpretable. Similarly, by giving up on amending our mental state ascriptions of others, we undermine our potential for understanding them: "To the extent that we fail to discover a coherent and plausible pattern in the attitudes and actions of others we simply forego the chance of treating them as persons" (ibid.: 221 f.). Also, a being we would find grossly uninterpretable is rather *arational* than irrational, meaning that our principles of rationality simply do not apply. The adjective "irrational" is better applied to single cases against a backdrop of understandable, rational attitudes: That is, if I am a generally relatable and interpretable person, I may occasionally do irrational things, like checking under my bed before going to sleep.

Consequently we should distinguish between two different uses of the term "rational": We may say that if Anne were rational, she would vote for the Green Party, and mean that for her actions to reach a maximum of consistency with her mental states (such as her political beliefs about environmental policies), she should vote for the Green Party. This falls squarely into the domain of mental ascriptions, for we would strive to find a way to make her voting behaviour consistent with her beliefs (and perhaps we would try to convince her to vote in what we perceive as a consistent manner). On the other hand, we might say the same thing, namely that if Anne were rational, she would vote for the Green Party, and mean something entirely different by it: namely that if her actions were to be most consistent with what we perceive to be the state of the world, she should vote Green. We might perceive the state of the world to be one of threatening global warming and waning natural ressources, and consequently think it only right to vote Green – even if none of Anne's beliefs are actually consistent with voting Green (as she might steadfastly deny global warming and the waning of natural ressources). *This* use of "rational" does *not* fall into the domain of psychological attribution, but into the domain of ethics, namely of determining what values should guide our actions: what's good and what's bad.[46]

For example, we can find Bertrand Russell lament: "Man is a rational animal — so at least I have been told. Throughout a long life, I have looked diligently for evidence in favor of this statement, but so far I have not had the good fortune to come across it, though I have searched in many countries spread over three continents" (Russell 1950: 71). Here, Russell does not mean to imply that human beings generally fail to weep over a loved one's death, fail to value acts of kindness, fail to eat when hungry and presented with food, or fail to understand why insurgents would not freely surrender their children to the oppressor. Rather, he means to criticise failures of critical and long-term thinking and the like: a lack of adherence to what he sees as reasonable values. Such criticism is warranted, but does not touch the fact that we have to assume a basic form of rationality in order to render someone's actions and mental states intelligible. I have sometimes found people to reject Davidson's claims on the grounds of such laments – but such rejections rest on an utter confusion of what form of rationality Davidson is talking about.

[46] There is also a tendency in recent moral psychology to want to substitute reason with an assumed process of deliberation which supposedly causes (and thus precedes) decisions, judgments or, generally, actions. However, since what we can find as usually preceding actions seldomly deserves the name "reasoning", but rather appears as an emotional impulse or an intuitive, automatic process, this operationalisation of reason has led to either of the popular claims that reason is not a cause of action at all, that humans are not reasonable beings, or that no such thing as reason actually exists. Jumping to such conclusions stems from confusing reasoning – a process which aims to conform to a normative ideal of what's reasonable – with a descriptive notion of an individual psychological process (compare Sauer 2012).

The application of the principle of charity with its sub-principles is both necessary and justified, since it is

> "not an option, but a condition of having a workable theory, [and thus] it is meaningless to suggest that we might fall into massive error by endorsing it. Until we have successfully established a systematic correlation of sentences held true with sentences held true, there are no mistakes to make. (…) If we can produce a theory that reconciles charity and the formal conditions for a theory, we have done all that could be done to ensure communication. Nothing more is possible, and nothing more is needed" (Davidson 2001a: 197).

In this way, we are making "maximum sense of the words and thoughts of others" (ibid.). Maximising agreement between interpreter and speaker is justified since "each interpretation and attribution of attitude is a move within a holistic theory, a theory necessarily governed by concern for consistency and general coherence with the truth" (ibid.: 154). This is why "[g]lobal confusion, like universal mistake, is unthinkable, not because imagination boggles, but because too much confusion leaves nothing to be confused about and massive error erodes the background of true belief against which alone failure can be construed" (Davidson 1980: 221).

Davidson's reliance on observability, behavioural evidence and the interdependence of subjective, intersubjective and objective knowledge entails that private states alone cannot constitute a basis for intentionality and semantic properties: "Perhaps someone (not Quine) will be tempted to say, 'But at least the speaker knows what he's referring to'. One should stand firm against this thought. The semantic features of language are public features. What no one can, in the nature of the case, figure out from the totality of the relevant evidence cannot be part of meaning" (Davidson 2001a: 235). Here, Davidson does not deny the existence of private mental states, but he firmly denies that they can figure (directly) into a theory of interpretation. "The crucial point on which I am with Quine might be put: all the evidence for or against a theory of truth (interpretation, translation) comes in the form of facts about what events or situations in the world cause, or would cause, speakers to assent to, or dissent from, each sentence in the speaker's repertoire" (Davidson 2001a: 230).

While the private subjectivist view is thus rejected, it should be just as clear that, on the other hand, nothing about this picture implies radical behaviourism: The statement "she was angry all the time, yet she never showed it" does not come out as contradictory; and the fact that someone displays calm behaviour at all times does not logically imply that she has never been angry. Rather, the idea is that there is a general connection between angriness-

behaviour and being angry, and that this connection generally justifies attributions of anger. Other mental states are justified analogously.

I.7.5. Full-fledged versus Attenuated Intentional States

Davidson famously held that for any belief to qualify as being about a tree, the belief's bearer must have "many general beliefs about trees: that they are growing things, that they have leaves or needles, that they burn. There is no fixed list of things someone with the concept of a tree must believe, but without many general beliefs, there would be no reason to identify a belief as a belief about a tree, much less an oak tree" (Davidson 2001b: 98, also compare Searle 2000: 107). Yet, there plausibly are weaker construals available of what is required in order to have meaningful thoughts. We should want to say that mice may fear snakes, for instance. Mice could not mean snakes by way of reference, since they have no symbolic system to rely on – and even if they had, nothing having to do with their being afraid qualifies as a symbolic action. They also cannot have the same full-fledged concepts of snakes that we do: If we are able to use the concept "snake" then because we have learned it. No such thing would be required of mice for them to be able to fear snakes. That is, if we are to attribute some concept of snakes to mice, then it would be a quite sparse concept, relating sensory sensitivity to certain stimuli associated with snakes to fearful behaviour. On the human side, invoking symbolic reference has a huge explanatory benefit when explaining the mind, and the fact that we can have certain thoughts which mice cannot have depends on our having developed concepts and symbolic reference culturally and on having grown up in an environment which is in large parts characterised by forms of symbolic reference. Animals may not lack meaningful mental states tout court, but they certainly lack *some* that are at our disposal, and our ascription of mental states to animals usually marks this difference, attributing attenuated concepts to them at best. There is a controversial and ongoing debate about to what extent and how exactly we should ascribe concepts and mental content to animals, which I will not indulge in here (see Jamieson 2009 for an overview); however, it should be clear that human psychology is distinctive insofar it is at least to a degree shaped and formed by our access to symbolic communication. Beliefs about quarks and hopes of rising stock markets would constitute prime examples (cf. Deonna & Teroni 2012: 23 f.), whereas we might fear snakes in a similar way as mice do. Yet, even in the latter case, we are able to develop a mental state which is impossible for mice to have; for example, an ophidiophobic herpetologist could certainly have a more complicated fear of snakes than mice

do. So, human psychological make-up ranges from areas which are not necessarily characterised by high-level intentionality to others which are.

The distinction between high-level and low-level capacities reflects the fact that some of the propositions we can have attitudes towards require higher cognitive capacities than others. For instance, in order to be able to fear stock market crashes or to be enticed by the beauty of a mathematical proof, we need to be able to have an attitude toward abstract objects – mice would certainly have a hard time sharing these mental states. In other cases, all that is required of anyone to have a certain mental state is that they show consistent behaviour as a consequence of certain environmental cues, where the cues are describable in cognitively sparse terms (such as fearful behaviour when snakes are around). For example, Daniel Dennett's deeply pragmatical view accommodates ascribing the intention to make a chess move to Deep Blue and the belief that the room is too cold to thermostats (cf. Dennett 1987: 22 ff.). The ascribed state may be "an attenuated sort of belief" (Dennett 2007: 87) and merely ascribe "hemi-semi-demi-proto-quasi-pseudo intentionality" (ibid.: 88) in contrast to full-fledged intentionality: "Just as a young child can *sort of* believe that her daddy is a doctor (without full comprehension of what a daddy or a doctor is), so a robot – or some part of a person's brain – can *sort of* believe that there is an open door a few feet ahead, or that something is amiss over there to the right, and so forth" (ibid.: 87 f.).

Without settling the matter how intentionality itself can come in degrees, we should acknowledge that Dennett's motivation to ascribe intentional states even when the requirements for the full-fledged attribution are not met is that it may just turn out to be exceedingly practical: The "intentional stance (…) pays off handsomely, generating hypotheses to test, articulating theories, analysing distressingly complex phenomena into their more comprehensible parts, and so forth" (ibid.: 87) and it can be adopted when faced with similar abstract functional structures between persons and person-like systems (compare ibid.: 89). As he goes on to say:

> "For years I have defended [the attribution of intentional states] (...) in characterising complex systems ranging from chess-playing computers to thermostats and in characterising the brain's subsystems at many levels. The idea is that, when we engineer a complex system (or reverse engineer a biological system like a person or a person's brain), we can make progress by breaking down the whole wonderful person into subpersons of sorts[,] agentlike systems that have *part* of the prowess of a person, and then these homunculi can be broken down further into still simpler, less personlike agents, and so forth – a finite, not infinite regress that

> bottoms out when we reach agents so stupid that they can be replaced by a machine" (ibid.: 88).

Suffice to say, the criteria for ascribing these attenuated intentional states can be considerably more permissive than requiring being capable of having attitudes toward, say, abstract objects. At first sight, a view like this contrasts with one like Davidson's. Obviously, children and robots often do not satisfy his holistic requirements. Still, it is important to note that Dennett does in fact accept the requirements Davidson voices – namely, as requirements for *full-fledged* intentional states. That is, we can stick with strict Davidsonian requirements, while allowing a more permissible pragmatic application of Dennett's intentional stance in other cases, ascribing *hemi-semi-demi-proto-quasi-pseudo intentionality*. We should simply pay attention not to confuse the two versions of intentional ascription, since higher-level cognitive skills may not be shared by non-persons such as animals, robots or brain parts, or at least they do not presently figure into our criteria for ascribing mental states to any of these (yet, we could imagine this fact changing with the development of new technology or with a change in our granting rights of personhood to animals, and it will also depend on insights concerning the cognitive capacities of the subjects in question).

I.8. Narrow vs Broad Content

I.8.1. The Central Issue

What does mental content depend on? Traditionally, two claims are distinguished: That the content of mental states either depends on properties intrinsic to those who have the respective mental state, or that it depends on something extrinsic to these, such as the individual's environment (and a third position may hold that it depends on both). Now, we know that the truth of the belief *that there is a glass full of water on my desk* trivially depends on the environment, namely on whether there actually *is* a glass full of water on my desk. But does the fact that my belief is *about* a glass full of water on my desk depend on the environment? Or, generally: Does any mental state's content depend on the environment? (See section I.3.) In a nutshell, to claim that mental content is *broad* is say that it does depend (inclusively or exclusively) on the environment, and to say that it is *narrow* is to say that it does not. I am going to argue for the view that mental content is broad in an inclusive sense.

(Broad content is also sometimes interchangeably called "wide content", but I will only use the former term.)

An intuitive assumption, which is likely to be held by many whose mindset has not been swayed by recent analytic philosophy, is that I can have any belief completely independently of what is going on beyond my own mind. We know that we can dream up or hallucinate all sorts of things – things which can at times have no representational relation to our actual environment at all. Historically, this assumption was most famously expressed by Descartes: in his *Meditationes de prima philosophia*, he explored a form of radical doubt in our beliefs, taking it to be plausible that the world could be entirely different from what we perceive and/or believe it to be. In fact, he took a methodically skeptic approach as a basis for reconstructing our entire epistemic inventory (i.e. the sum of our knowledge), and thus as one basic form of methodology for philosophy (see Descartes 1965: II., 7-10). However, I suggest we should primarily take his as an epistemic point: If at any given time, our perceptions may deceive us, then how can we actually *know* whether our beliefs are grounded in reality? (Compare section I.7.)

But the matter becomes far more enigmatic when it comes to questions of content, which Descartes did not address directly: While a systematic sensory access to the world is certainly a basis for knowledge, it is not all that clear how mental *content* is formed depending on sensory input, and what difference exactly a potential independence of sensory input from reality would make to semantic matters. Eventually, it very much depends on what you take the nature of mental content to be. For example, if you happen to believe that concepts are God-given Platonic ideas in the mind, and forming true beliefs about them amounts to the enterprise of somehow matching them with what is real, then whether sensory perception is independent from how the world actually is may impact your epistemic faith (in the truth of your beliefs), but it has no bearing on mental content at all, since content is merely a matter of forming beliefs out of these God-given ideas which have no traction with reality anyway. Given such a view, one could have the belief that *there is a full glass on my desk* no matter whether there actually is a glass on my desk, or whether I have actually ever been in the presence of one, or whether any glasses and/or desks have actually ever existed. All that would matter is that such ideas can exist in the mind.

Then again, why should we be satisfied with a view which does not even begin to explain how anyone can arrive at such a belief, true or not? Ultimately, the question we want answered is this: What facts about me and/or the world decide what my mental states are about? (And I take it that "they simply *are given* to us this way" is not a satisfying answer.)

On the one hand, we certainly require specific cognitive abilities in order to have meaningful mental states. That is, we have to be able to know, at least in principle, what a glass of water is to have a belief about it, and we have to be at least in principle able to use the associated concept somewhat properly (which, on a weaker reading, does not need to amount to linguistic capacities, but instrumental knowledge, such as knowing what to use a glass for, and connecting it to related concepts such as transparency). On the other hand, these capacities are developed in interaction with certain external matters of fact, if only because I have to learn many of them in interaction with my environment. So, in order to employ them properly, I need to be cognitively able to learn the relevant concepts, which points to internal facts, and I need to actually learn them, which points to external facts. Hence, an interaction between cognitive "internal" and environmental "external" facts should be assumed.

Elaborating on these matters in detail is what is necessary to address the question whether mental content is narrow or broad. As I have mentioned, I do not believe the Cartesian view to help us iron these out, although what we should note is that the mere fact that he believed in the possibility of all our beliefs being wrong implies assuming independence of content from environment. That is, if our senses have been systematically deceiving us, then we must have acquired mental content in a way which is independent from our sensory connection to the world. Still, this mere assumption won't lead us far.

In opposition to Cartesian skepticism, Quine and Davidson took mental content to have a firm basis in what the world is like (see section I.7.4). The question what an utterance is about is ultimately settled by checking whether the utterance has a systematic connection to certain environmental (i.e. objectively assessable) stimuli (see Quine 1960: chapter 1). And mental content is essentially reconstructed from assertions, thus dependent on the meaning and truth of utterances (see e.g. Davidson 1980: 256 f.). It is no surprise that Quine and Davidson have fashioned this view into a thorough rebuttal of the Cartesian skeptic: The wrongness of all our beliefs at once is entirely inconceivable, since the content of these very beliefs depends on what the world is like; and beliefs are ascribed by creating a maximum of correspondence between what is true and what is believed (cf. ibid.). In fact, we should expect all theories which take content to be intimately connected to things or events in the world to be more or less suited to oppose Cartesian views, and to not allow us to conceive of the world as being radically different from what we believe. I will elaborate on one other example, namely Putnam's causal theory of meaning as underpinning an argument for broad content, in more detail in section I.8.3.

While I have now only sketched two radical alternatives – the mind's directly given access to content versus taking content as the operative aspect of empirical psychosemantic theories –, ultimately, the question whether mental content is broad or narrow will decisively be settled by whatever picture we lean towards. Thus, the ultimate goal of this section consists in my clearing up what exactly follows from my own theory about the content of mental states, and how it follows. But first: some preliminaries.

I.8.2. Dependency, Intrinsic and Extrinsic Properties

First off, to say that mental content *depends* on anything can mean a number of different things. The weakest explication of this dependency is that the properties in question have some effect in bringing about the mental content in question. To use an analogy, any particular sound depends on the space it occurs in, yet the space is only one factor in how the sound turns out. Without knowing some of the sound's other properties, merely knowing about the space it occurs in will not be very helpful in learning what the sound is like. Analogously, some believe that some mental states can only be had by beings who speak a language (see section I.7.5), and in this sense mental content would depend on having learned a language. Yet, merely knowing that a particular person has learned any language will not be enough to learn the exact content of their thoughts.

On the other end of the spectrum, the strongest explication of this dependency is to hold that mental content is completely determined by the respective properties (and perhaps derivable from and/or reducible to these).[47] To say that it is determined can once again mean a number of things: for example, if a set of properties {P} determines content C, it can mean that it does so by (natural) law, and that there is a causal relationship between {P} and C, which may enable us to derive knowledge about C from knowledge about {P}. The latter explication of dependency is stronger than the former because you can know about all sorts of factors which are required to bring a certain phenomenon about and still not derive this phenomenon from said factors (such as the quality of a sound from the space it occurs in, or

[47] In this paragraph, I included both ontological and epistemological ways of stating the respective dependency relations, even though these may be held independently of one another. That is, it may be the case that, ontologically, A is completely reducible to B, without A's ever actually being derived from B. Of course, the case becomes more complicated for abstract objects: If A and B are abstract entities (which I take mental states to be), then what other reason could we have for judging A to be reducible to B other than A's being derivable from B (i.e. that our knowledge about A depends on nothing but our knowledge about B)? Consequently, I am leaning towards epistemic formulations, and have been concentrating on these a bit more than on ontological ones – a trend which is to explicitly continue in this section.

the content of a thought from someone's spoken languages). Thus, I take the relevant claims about dependency relations to be situated in a range between the weaker "being necessary for" and the stronger "being sufficient for": "C depends on {P}" can mean anything between "{P} is necessary for (there being) C" and "{P} is sufficient for C" (or, as an epistemic alternative: "C is derivable from {P}").

Secondly, while I have introduced the distinction between broad and narrow content in terms of dependency on the environment, you may want to ask what exactly mental content depends on if it does *not* depend on the environment, and how to characterise the distinction between the mental state's bearer and her environment. An ad hoc answer to both these questions consists in pointing toward the distinction between intrinsic and extrinsic properties: Narrow mental content only depends on intrinsic properties of the individual, whereas extrinsic properties are had in virtue of an individual's place in her environment. For example, my currently being warm is due to my sitting next to a heater, and I assume most will thusly want to construe it as an extrinsic property of mine. However, there are numerous relationships between this extrinsic property and my intrinsic properties: Maybe you believe that the properties of all molecules I am made up of make for nicely intrinsic properties. Yet, their movement is undoubtedly caused or at least crucially influenced by my sitting next to a radiator – so shouldn't it rather count as an extrinsic property? Undoubtedly, things which are intrinsic to myself, such as the molecules I am made of, will have many of their properties due to extrinsic factors. If we were to go through a list of my properties, I assume many will not uncontroversially fit into either the category labelled "intrinsic" or the one labelled "extrinsic" (see also Lewis 1983a).

So, if we are to thoroughly build on the instrinsic/extrinsic distinction, our first aim should be to give it a more solid footing. I propose that what we can do is exploit the conceptual proximity of the intrinsic/ extrinsic distinction to the "internal/external" distinction on the one hand and to the "essential/non-essential" distinction on the other. The former, at least if taken literal in the way we usually do in everyday talk, is close to the individual/environment distinction, since both are usually characterised and informed by organismic concepts: by concepts of our physical bodies, of what's beneath our skin, and so on. (And it is no surprise that Putnam's respective slogan reads "'meanings' just ain't in the *head*!" [Putnam 1975: 227] – see my discussion of his view in the following section.) A distinction based in these concepts can occasionally get fuzzy: When we are dealing with technological implants, when our minds are in some form controlled externally (through "brainwashing" of any form), when we expand our minds into technology (compare Chalmers & Clark 1998), or the like.

Still, this is probably the most familiar and intuitive distinction at play here, and we should expect especially the notion of our environment to be informed by notions of what's beyond our body.

Current research in the cognitive sciences has a lot to say about the psychological mechanisms underlying our phenomenal body image (see e.g. Botvinick & Cohen 1998). However, once we go beyond phenomenology – beyond what *feels* like it is our body, and what doesn't –, it is hard to find anything inside our body which is in a strong sense independent of the environment. We will at least find that every part of our body *is* part of our body *because* of the environment – because the genetic blueprint for our body has evolved in interaction with a specific environment, because our body is built and maintained by nourishment (i.e. literally taking in parts of the environment), and so on, down to said molecular movement. Thus, our criterion for judging whether anything is intrinsic will have to be considerably weaker than demanding a complete (causal, explanatory or conceptual) disconnectedness from the environment if it is to be applicable at all. Such a criterion is far easier to come by for a biological concept than for a purely physical one. Thus, I will define what counts as intrinsic as follows: a physical object inside our body is intrinsic if and only if its being inside the body can be traced back to organismic functions (i.e. if it either serves an organismic purpose, or because it happens to be inside us because of bodily functions). Again, I need to stress that what underlies this criterion is not a purely physical notion, but one of biological organismic function. Only insofar an organism makes use of physical objects and processes are these graspable as distinct (or in this sense "independent") from the environment, and hence as intrinsic.

However, what about properties we usually take to belong to ourselves, to characterise us and define us as who we are – such as our social and/or judicial status, and many psychological characteristics such as our character traits –, properties we tend to think of as intrinsic and essential to our personal identity, which are clearly not literally "inside" us, i.e. inside our bodies? Bodily changes are often connected to other personal changes: Bob Hoskins ceased to be an actor because of Parkinson's disease; but in a very important sense, during its progression he did not cease to be Bob Hoskins, or the person who starred in "*Who Framed Roger Rabbit?"*. In fact, no conceivable change in his nervous system would have been able to change that. And most citizens will remain citizens for all their lives, no matter the manifold bodily changes taking place between their births and deaths. Apparently, many defining aspects of our identity are entirely unaffected by bodily changes. This is because the criteria by which we judge someone to remain the same often depend on social institutions, on

social relations, and all sorts of things which may not themselves be fully describable as intrinsic properties of the person, and for whose establishment and maintenance bodily changes are not decisive, much less anything physical making up the body.

Many properties we tend to think of as essential to us will consequently not be intrinsic in the biological sense specified above, and vice versa. For example, the fact that I am 1,90m tall might not go a long way to clearing up questions about my identity as a person, but it is a property I have, and it also is intrinsic to me right now. Granted, measuring someone's height does depend on the environment in many ways: on the institutionalised scale of measurement, on a certain stage of scientific progress, on the sheer possibility of using an external object in order to measure myself – what else does measuring something mean beside putting it into a definite relation to the environment? Still, we can certainly distinguish between the preconditions for measuring and the property which is measured. And all I have said so far about a dependency on the environment only applies to measuring my height, not to my *having* the respective property of being 1,90m tall. Thus, it should still count as intrinsic.[48]

I.8.3. Two classic arguments for broad content

Given said difficulty of coming up with intrinsic properties which go beyond physical properties in relation to organismic functional terms, it comes with no great surprise that the most famous argument for broad content was delivered in terms of physical properties of bodies. Specifically, Hilary Putnam asked us to imagine two "physical twins"; that is, two

[48] On a side note, it just so happens that the property of being 1,90m tall is identical to the property of being ten times as tall as the wavelength of the GPS L-band frequency. Being ten times as tall as the wavelength of the GPS L-band frequency certainly is an extrinsic property, since it depends on the environment in numerous ways (on the properties of the respective frequency, on its being used for GPS, and so on). Also, my being 1,90m tall implies that I am considerably smaller than the Eiffel Tower – again, an extrinsic property. Still, my height is only coincidentally identical to ten times the wavelength of the GPS L-band frequency, and only coincidentally implies my being smaller than the Eiffel Tower. That is: If my environment were significantly different – if the GPS wavelength were different, and if the Eiffel Tower only existed as a miniature, or not at all – then I would not have these properties (despite my still being 1,90m tall). Which is not only the same as saying that these properties are extrinsic themselves, but also that their being related to (i.e. their being identical with, or implied by) my intrinsic property of being 1,90m tall is extrinsic.
This problem can be solved twofold: by either introducing modality and insisting that the supposedly intrinsic property, namely my height, which also happens to be describable in terms of many extrinsic properties, should be characterised as a property which is necessarily intrinsic (i.e. for which there is no property which is identical with it *and* which is intrinsic in all possible worlds). Or we could make being intrinsic relative to descriptions: A property is intrinsic iff it is intrinsic under one or more possible true descriptions. For example, my currently being in Munich is not intrinsic under any description, and thus extrinsic. My being 1,90m tall is intrinsic under at least the description that I am 1,90m tall, and can thus count as intrinsic, even though there are many alternative descriptions available, such as that I am ten times as tall as the wavelength of the GPS L-band frequency.

people whose physical states and histories are exactly alike, but who live on different planets (cf. Putnam 1975). In this thought experiment the twins' respective environments are also alike (at least in terms of the effects they have on the twins' bodies), except that what is water in our setting is replaced in the other, which Putnam calls "twin earth", with a liquid which shares its macroscopic properties with water. However, its molecular structure is not H_2O, but XYZ – a fact which the twins have no access to (much as it was until the late 18th century, before Henry Cavendish discovered the chemical composition of water). Thus, they developed the linguistic dispositions underlying their intentional properties based on identical observations, making both competent users of the concept "water" (or, depending on what you take to be necessary to make them competent speakers, perhaps the totality of observed properties in a crucial part of the linguistic community). For example, they would both readily assert that water is usually wet, transparent, and so on. Now, given that in this example, the molecular structure of water had no hand in shaping their dispositions, we should intuit that one such twin's thoughts about water in the H_2O setting do indeed mean water, whereas the other twin's thoughts mean XYZ. This is because the term "water" refers to a natural kind, something whose underlying structure explains its observable phenomena, such as wetness (also see Kripke 1980 and section I.4.3). And matters of the underlying structure are matters external to the twins' thoughts.[49] So, the bottom line of Putnam's argument is this: Whenever we are thinking about natural kinds, whose essential properties we may not always be aware of, and knowing which is not necessary in order for us to be able to refer to them, the content of our thoughts depends on what the physical world is like.

But perhaps we can even say something about mental content which is not about natural kinds? Tyler Burge has done just that with his argument for "anti-individualism" (see Burge 2007: Introduction). According to his argument for broad mental content, which is also known as "semantic deference", we can competently use certain concepts, even while deferring exact knowledge about them to experts in the respective field. On this view, the referents of some of the concepts I am thinking about depend on my environment insofar as they depend on what the experts have to say about them. For example, we might concede that a layperson can believe that she has arthritis without having exact medical knowledge about arthritis. Thus, the content of the layperson's belief depends on her social environment.

[49] The objection has been voiced that the two cannot be physical twins, since the one's body in the water-setting would be composed of more than 70% water, whereas the other would in its place be composed of XYZ. Since the molecular structures are different, the two bodies cannot be physically identical. Putnam's argument should in principle be salvageable though – how about we swap water for mercury or the like?

If we accept both arguments, then we should accept that mental content is broad if the respective mental states are about natural kinds or about terms whose referents are decided by experts. Still, this leaves us asking about a lot of other potential referents, and whether the content of my thoughts about these, too, depends on the environment. In what follows, I will take a more general approach to mental content, and I will argue that mental content, by its very nature, is broad.

I.8.4. The Derivability Argument for Broad Content

As we have seen, what we take to be constitutive of or essential to someone can easily go beyond what is literally inside their bodies. In many cases, this is plain to see, as in the case of social heritage or relationship-status. In others, such as those pertaining to mental properties, it is controversial. But the claim that mental content is narrow apparently clashes with our methodology of relating internal properties of agents to the mental content they have. That is, in reconstructing, inferring or deriving mental content from internal properties (or in "reducing" it to internal properties), we cannot help but take the environment into account. The narrow content claim in terms of derivability means that mental content needs to be reconstructible from internal properties only. If this cannot be done, then mental content at least epistemically depends on environmental factors.

For any B to be derivable from A a relationship between A and B is required which supports this derivation. We can describe this relation as a function "f(A) = B", specifying which B can be derived from which A. If nothing else enters into A but internal properties of an agent, this means that the derived mental state B is independent from the environment – that all possible external factors are irrelevant. So, it seems that for the narrow content claim to be true, every property from which a mental property is derived, inferred or reconstructed has to be intrinsic (best understood in the organismic sense, see I.8.2).

However, for the mental state to completely depend on intrinsic properties, it is not only necessary that said function does not require taking anything external into account, but that it *never* does. That is, in order to practically establish a derivation function [f(A) = B] from intrinsic property to mental content, we need to first find out how *A*s systematically relate to *B*s. If the exclusion of environmental factors were a methodological requirement for establishing relations between intrinsic and mental properties in the area of cognition, it would surely usually be violated. For example, while we may be able to infer a particular

toad's ability to perceive prey from the toad's internal make-up, it is the sheer fact *that the toad's ability to perceive prey depends on the environment* which enables us to trace this ability to its internal make-up in the first place (cf. Ewert et al. 1996 and my section II.6). For in order to establish mental content dependency on internal features, the cognitive function of an internal feature has to be identified as such, and this is tied to identifying the proper environmental "aim", i.e. the proper input/output relation of the underlying (neural) mechanism (cf. Sullivan forthcoming). That is, we only know whether the function of a given neural mechanism is related to prey-capture if we can trace its activity to predatorial behaviour, and predatorial behaviour is characterised in relation to behaviour aimed at actual prey (i.e. the predator's environment).

The notion of cognitive function is also tied to the environment in a second way, namely in conceiving the mechanism as being evolutionarily selected in an environment in which it has proven advantageous. Compare Millikan's and Neander's teleofunctional accounts: "A proper function of (…) an organ or behaviour is, roughly, a function that its ancestors have performed that has helped account for the proliferation of the genes responsible for it, hence helped account for its own existence" (Millikan 1993: 14), and "some effect (Z) is the proper function of some trait (X) in organism (O) iff the genotype responsible for X was selected for doing Z because doing Z was adaptive for O's ancestors" (Neander 1995: 111).[50] Again, what this tells us is that cognitive directedness is only ever properly construed as such by taking an agent's (or its species') environment into account.

So, our knowledge about an agent's interaction with the environment plays a crucial role in establishing this relation between its internal (cognitive) features and their output as functions. Thus, the latter will not at all be reconstructed *purely* from internal properties, but rather, the external properties become background conditions for the function(s) which allow(s) us to infer one from the other. So, the sketched argument for narrow content from derivability must fail, since the demand that the relevant derivability function takes into account nothing but internal features will never be satisfied. Since our methodology of understanding cognitive mechanisms always requires environmental knowledge (since the environment is both a conceptual part of cognitive functional characteristics as well as teleosemantics), the mental function associated with this mechanism cannot be described in

[50] These accounts are called "teleofunctional" because they tie a mechanism's specific function to its inherent functional "aim" (Greek *telos, τέλος*). This notion does not require that evolution itself is directed towards an aim, but rather that the functionality of biological traits is tied to evolutionary explanation (compare Thompson 2008: 78 f.). For an overview of teleofunctional accounts see Hazlett 2013: 182 ff..

terms of narrow mental content – just the contrary. For this reason we should think of it as broad.

So far, I have argued against claiming that apparent cases of mere derivability of mental content from internal bodily (i.e. organismically "intrinsic") features are sufficient for claiming that the content in question is narrow. What has to be ensured is that the principles underlying this derivation, namely the relation between internal features and mental content, have been arrived at without taking external features into account – and this cannot be achieved in principle. In the cited case of reconstructing a toad's behaviour from its internal features, the point should be clear: we know about the toad's cognitive function because we have observed its predatorial behaviour in relation to its internal features (i.e. the implementation of the cognitive mechanism in the toad's nervous system). If we hadn't observed the predatorial behaviour – which is a feature of the toad *within its environment* – we hadn't arrived at the necessary principles. The analogy between such cognitive capacities and psychological attitudes, roughly sketched, is this: The fact that we can have mental content rests on our having certain cognitive functions, and if we had principles of deriving content from features of functional implementation, we would only have these because environmental features would have entered into them. This point will also be further explored in my discussion of "mindreading" experiments in chapter II.4.

I.8.5. The Constructionist Argument for Broad Content

In what follows, I will show how a dependency of mental content on the environment follows from my own account of mental states. It crucially rests on the constructionist idea that mental states are real objects in virtue of being explanatorily valuable concepts in psychological theories. In this concluding section to my discussion of narrow versus broad content, I will place my basic constructionist premises into a larger context of explanation in psychology versus explanation in terms of cognitive mechanisms and especially exploit the notion that reliance on semantic content is a characteristic of intentional psychological explanation, while the latter is a form of causal explanation.[51]

It has been objected by Fodor (1989 and 1991), that whatever distinguishes broad from narrow content is not causally efficacious. Thus, the type of content central to psychological

[51] In the literature, we can occasionally find an antagonism between semantic and causal explanation (cf. e.g. Smolensky 1988). On my account, intentional psychological explanation is both semantic and causal. However, it should be understood that there are forms of explanation which are either only semantic or only causal.

theories cannot be broad, since what is important to psychological theories is that they explain behaviour in terms of being caused by mental content (the basic idea being that our beliefs, desires and intentions causally explain our actions). Fodor originally supplied an individualistic analysis of causal powers (in his 1989: chapter 2), which holds that causal powers of an individual are rooted in their intrinsic properties. One response could consist in outrightly rejecting this analysis of causal powers (see Rechenauer 1994: 75) and argue for the position that "causally relevant mental states, which are implicitly or explicitly invoked in normal practice and scientific theory, are individuated externally" (my own translation from German of ibid.: 71). I will indeed do just this; however, it will not crucially rest on any specific analysis of causal powers as rooted in either intrinsic or extrinsic properties, and thus, I feel that discussing Fodor's analysis at length would only unnecessarily complicate things here.[52] Ultimately, I aim to show how to reconcile a picture of intentional psychology as a form of causal explanation with the insistence that the content invoked in these explanations is broad, and I hope that by the end of this chapter it will be clear how an account such as Fodor's relates to my own.

Firstly, mental states are attributed on the basis of observable evidence. This need not exclusively be behaviour, but it undoubtedly plays an important role. For example, looking through my drawer may lead you to attribute both a certain degree of absent-mindedness and compensatory practicality to me. A friend's report may lead you to do the same. In neither case did you observe my behaviour, but in both cases, the connections to my behaviour are evident. In fact, it seems very hard to understand being absent-minded or practical not as at least implying a certain regularity in specific kinds of behaviour. Thus, in turn it is reasonable to assume that the evidence used for any such ascription relies at least in part on behaviour.

Secondly, we know that behaviour is crucially tied to neural coding. As Fodor rightly points out, it is "preposterous to suggest that neurological (or biochemical; or molecular) states should be taxonomised by reference to the sorts of properties that distinguish [physical] twins" (Fodor 1991: 11, fn. 9). Here, going back to Putnam's thought experiment I just sketched in section I.8.3, the idea of physical twins means two persons who share all physical properties, and Putnam invoked these to show that the content of their mental state can nonetheless differ. What Fodor calls preposterous is the notion that anything the neurosciences could find out about these two physically identical "twins" could differ

[52] For a better impression of Fodor's analysis, see his 1991: 9 (for his "Schema F") and ibid.: 24, where he goes on to claim that "for the difference between being [one causally relevant property and being another] to be a difference of causal powers, it must at least be that the effects of being [one] differ from the effects of being [the other]. But, I claim, it is further required that this difference between the effects be nonconceptually related to the difference between the causes".

between them, i.e. the suggestion that neurological, biochemical or molecular taxonomies should have to refer to anything which could potentially differ between two physically identical twins. (And on my view, this is really just to say that neurobiological statements do not invoke semantic/intentional content.) Consequently, if sameness of physical properties implies sameness of intrinsic properties, then so should sameness of "neurological (or biochemical; or molecular) states".

Thus, there is at least the following way of relating behaviour to intrinsic properties: A certain stimulation of the senses leads to a neurological coding which in turn leads to a behavioural output. There are some caveats in this description; for example, senses can be stimulated without outward behaviour to occur. The important point, however, is that any stimulation of an organism's senses causes a change in its neural networks, and that its motor behaviour has to be completely explainable in reference to the causally relevant state of the connected neural network (namely, by way of its nerves eventually innervating and stimulating the muscles responsible for movement). So, far from insisting on a complete derivation of behavioural output from sensory input, the importance lies in methodologically assuming that motor behaviour is explainable by nothing but neural activity, which in turn is at least modulated by sensory stimuli. Any other determinants of the neural network, such as internal chemical changes, are construed as intrinsic properties as well. For example, the discharge of one neural transmitter at a specific place in the nervous system might cause the discharge of another neural transmitter at another place. Construing things this way, we arrive at a purely intrinsic description of behaviour: The stimulation of sensory organs is a physical state of the organism in question, just as the states of the neural network and its eventual motor output. If an organism's behaviour is in exactly this sense taken to be a purely internal disposition to react to certain stimuli in such-and-such a way, and if this is all there is to an account of the individual's causal powers, then of course these are intrinsic.

It should be noted, though, that the concept "behaviour" used in this context is ambiguous: I have tried to stress that what we are really after is "motor behaviour", i.e. an activity of the organism's muscles. If this is so, then we have failed to give any intentional weight to the behaviour in question: That is, we did not say whether the organism's movement is intentional – whether it is actually aimed at whatever has caused the sensory input, or related to it in a psychologically interesting way. Behavioural descriptions, as belonging to the proper inventory of intentional psychology, are couched in intentional vocabulary, at least in order to take an organism's systematic and/or functional relations to the environment into account. This applies even in the most sparse attributions of intentional states: That a toad aims to

catch a worm is to describe its behaviour as intentional in this sense; it is, of course, *not* intentional in the full-fledged intentional psychological sense that the toad needs to invoke any semantic content in order to catch the worm, but it is insofar as the description is not about a stimulus, but about a worm, and not about resulting motor action, but about a movement within the environment, with an inherent aim that lies beyond the toad's organismic boundaries.

Here, I will briefly have to come back to the conceptual complications which lay at the outset of my considerations regarding the distinction between narrow and broad content. This is because, even when trying to formulate an input-output relation along the lines I have just attempted in intrinsic terms, we cannot help but conceive of the causes of both sensory stimulation and the nervous system's origin, which is tied to its function, as lying outward: The perception of anything that is not just hallucinated obviously depends on the environment, and so does the evolution of the brain. We can conceive of any number of environmental changes which might either change our sensory stimuli or the evolution of our nervous system, and thus we cannot avoid as thinking of either as crucially depending on the environment. Yet, the physical state of Putnam's twins was certainly caused by the environment in this sense as well; and if we accept his notion of intrinsic properties, then I have just shown a way of relating these to behaviour. I feel that this is ultimately a definitorial choice, depending on how strongly or weakly we'd like a given feature's dependency on the environment to be in order for us to call it intrinsic or extrinsic. I have suggested an organismic notion as most favourable; what matters here is that the distinction between such a conception of intrinsic properties still significantly contrasts with a conception of mental content as *not* depending on intrinsic properties understood in this way. That is, even if we accept that our physical state, including the state of our brain, depends on the environment, there still is the possibility of distinguishing narrow from broad content: Narrow content depends on our physical properties, broad content depends on more than those, or on features which are not only caused by the environment, but which are necessarily described as depending on the environment (for the unwieldy details, see footnote 48). Remember Fodor's suggestion that neurological states should not be taxonomised in terms of the differences between physical twins: This implies that, if there is broad mental content, then at least some (environmental) properties which have the power to change this content would leave our physical and neurological properties unchanged. That is: If the notion of broad content is true, then it has to be conceivable that my being transported from one environment to another (such as Putnam's twin earth, see I.8.3) changes *only* the content of my belief without changing my

physical state. That my physical state also depends on the environment in some ways is an unfortunate byproduct of the fuzziness of the concept of dependence. To say that *a property is intrinsic iff it does not depend on the environment* is too vague for there to be anything that is fully intrinsic, even our physical states.

But here is where the fuzziness ends, and where we can highlight the characteristics of intentional mental states in order to sharply bring out the contrast between this special notion of intrinsic versus extrinsic properties (namely a notion which allows for physical properties of the body to be intrinsic). As pointed out in section I.7.5, Dennett permits the ascription of "an attenuated sort of belief" (Dennett 2007: 87) to a thermostat (cf. Dennett 1987: 22-34), such as the belief that the room is too cold. I take an ascription of "hemi-semi-demi-proto-quasi-pseudo intentionality" (ibid.: 88) along these lines to work because it captures two important things, namely that (1) the thermostat possesses (or is made up of) an internal mechanism which causally relates its sensor to its adjusting the heater and (2) a norm which describes a well-working thermostat (i.e. only a specific adjustment of heating relative to the sensor is adequate, otherwise the thermostat counts as being broken or maladjusted). This norm is responsible for a properly working thermostat's adjusting the heater in a specific way depending on its sensor. Of course, the thermostat itself need not know anything about or partake in anything that produces this norm. It cannot suffer from cold, and it cannot ascribe beliefs to its fellow thermostats about their feeling too cold, or to anyone else. And it doesn't need to, because the norm for a well-working thermostat is not determined by the thermostat itself but supplied externally – by its designers, manufacturers and installers, who can feel cold and who can hold beliefs about whether a group of friends in their living room is feeling cold and whether they should therefore turn up the heat (cf. Dretske 1988: 40, Dennett 1987: 33).

And that's the difference between thermostats and intentional agents: intentional agents don't have anyone to normatively "adjust" them beside other intentional agents. Thus, they not only have to possess the mechanism underlying the belief-ascription (i.e. the mechanism yielding the behavioural output which counts as evidence for the belief ascription given a sensory input, which we were indeed able to describe in terms of intrinsic properties of our nervous systems before; also see II.8.3), but the ascriptions themselves crucially require the norm which relates the mechanism to the environment. For example, explaining someone's turning on the heat by saying that she believed that it was too cold is not just to say that she felt too cold, but also that there is a psychological relation between feeling too cold and turning on the heat. If there were no relation of the latter kind, then anyone's feeling cold

could not explain their turning on the heater. This relation is not an intrinsic property of any intentional agent: it is a workable law of psychology. And given mental constructionism (see I.5), only when we have such laws is it possible and fruitful to ascribe content to ourselves and our peers. And it is exactly in this sense that the content of intentional states depends on the environment, and in which it is broad. It depends on the environment not necessarily in the sense that swapping environments would change one content for another (although we can envision these cases too, such as in Putnam's thought experiment), but in that without certain features of the environment, there would be no content at all.

This is why intentional states are not merely alternative descriptions of states of the intrinsic mechanism(s) underlying them, to be automatically replaced by these when we know all about them, and of whatever intrinsic causal powers these may have: they are at the same time providing information about the conditions under which this mechanism is adequate. That is, they incorporate both information about the agent's internal cognitive state as well as information about the environment.

To thoroughly drive this point home, I'll once again invoke the example of the toad, and how its cognitive mechanism underlying its predatorial behaviour depends on the environment: Toads have a cognitive mechanism which, on the level of sensory stimulus, operates on the perception of an elongate apparition. It can be shown that the toad shows stereotypical predatorial behaviour when confronted with stimuli which produce this perception (for the details see section II.6). Thus, we know that the toad uses a cognitive mechanism which takes a definite stimulus as an input and produces definite (ranges of) behaviour as an output. The stimulus, namely the elongate apparition, can be construed as the "intentional object" of the mechanism, in the sense that the mechanism "aims at" such stimuli. However, this whole story would be rather mysterious if we could not also assume that the toad's environment was such that this mechanism would be favourable for it. That is, its environment has to be such that the mechanism allows the toad to consistently catch prey when perceiving elongate things. However, in the physical make-up of the toad, there is no such thing as an additional cognitive encoding of elongate apparations *as* worms. Rather, if placed in an adequate environment, no such additional encoding is needed.[53]

That something internal represents something external is not an intrinsic fact. The toad's neural network, which we take to implement the respective cognitive mechanism, could not

[53] This point is adapted from "Simon's Ant" (see Haugeland 1995: 209). There, the idea is that the at times beautifully complex paths ants take are tracable back to rather simple cognitive algorithms, and the ensuing complexity is a result of the interaction between the cognitive mechanism making use of the algorithm and the environment.

possibly hope to encode "worms themselves" – for there is no such thing as an encoding of worms themselves. There are internal cognitive mechanisms, which, in favourable environments, produce behaviour which is aimed at worms. Just as it is in the case of human agents: The fact that we have intentional states is nothing which could be directly (or "intrinsically") encoded as part of our cognitive make-up. Rather, the fact that there are certain objects in the environment which we are aimed at in manifold and intricate ways is, once we are the subject of psychological theorising, described by systematically invoking intentional states – by interpreting the cognitive makeup of agents as standing not merely in causal relations to the world, but in largely rational ones (see I.7.4).

Still, there is a certain ambiguity in what we are really aimed at – in the toad's case, is the toad not really aimed at elongate objects rather than worms? That this is in fact not so is only ensured by the toad's environment, which produced the respective mechanism, and which ensured that there were enough worms making up for those elongate objects which were not worms, which, intentionally speaking, can "deceive" the toad. Consequently, I propose we'd best distinguish between a proximal "stimulus content" (namely all things the toad's mechanism can be "tricked" with) and a distal "functional content" (namely the mechanism's proper object). The latter being, if psychological theorising permits, the object of an intentional mental state, or, much the same, the content of a mental state. Narrow content thus equates to proximal content: The intrinsic properties of agents are such that under a given stimulus condition, they output a certain behaviour (and how this happens exactly is under investigation by the neurosciences). Broad content equates to distal content: The intrinsic properties of agents, which determine them to react in such-and-such a way to such-and-such stimuli, make sense, or are justified, and are ultimately described as intentional states, only when we add information about the environment.

Here, I am sympathetic to Ned Block's "mapping theory" (cf. Block 1991), which says that, given a particular environment, a narrow content is that which determines a particular broad content – that is, narrow content *maps* environments onto broad contents. (It should be clear that the relevant environment is not just the one the subject is currently in, but rather also the environment in which she acquired the relevant beliefs and other mental states.) Of course, I hasten to add, a particular environment only "determines" broad content by way of intentional methodology, namely interpretation. And interpretation, as we know from Quine and Davidson, does not yield definite content, but an indeterminate range of ascribable contents, with this range being constrained by principles of coherence and correspondence (see section I.7.4).

This notion of mental content being broad in the sense explicated above chimes with the view that intentionality is established in interaction and not specified individualistically. Rather, if an individual has certain cognitive dispositions to develop the foundation for intentional states, then she can learn to partake in the practice of mutually ascribing these. And we should expect this interaction to have specific effects on the development of our nervous system. Obviously, the claim that partaking in any practice changes our nervous system is trivial, since everything we experience changes our nervous system. The claim is rather a combination of two points. One is made by Quine in reference to the acquisition of linguistic capabilities: "Different persons growing up in the same language are like different bushes trimmed to take the shape of identical elephants. The anatomical details of twigs and branches will fulfill the elephantine form differently from bush to bush, but the overall outward results are alike" (Quine 1960: 8). Another is made by Buller in reference to the evolutionary advantage gained by our brain's plasticity:

> "According to our best evidence to date, the brain structures that perform specialised cognitive functions — and that would have been involved in generating cognitive solutions to adaptive problems throughout our species' evolutionary history — develop through a process of diffuse proliferation of brain cells and connections followed by a "pruning" that shapes this diffuse connectivity into relatively specialised structures. That is, functionally specialised brain structures are produced by a process consisting of both "additive" events (the formation and migration of brain cells and the formation of neural connections) and "subtractive" events (the pruning of synapses through cell death and axonal retraction) (Elman et al. 1996). In this process, gene-directed protein synthesis is involved in the additive events that build the diffuse connectivity with which brain development begins. The subtractive events, however, are not under genetic control. Rather, the subtractive events occur through cell competition, whereby cells with the strongest patterns of innervation (primarily from sensory inputs) retain their connections and the others die. Thus, genes specify the proteins involved in the additive events during brain development, but the forms and functions of brain structures are then shaped by environmental inputs. So the specialised brain structures we have are primarily environmentally induced, not "genetically specified"" (Buller 2006: 200 f.; also compare Selemon 2013).

Taken together, the picture is this: There is both a phylogenetic (invoking distal environmental influences having happened in our evolutionary history) and ontogenetic story (invoking proximal environmental influences happening in our lifetimes) to tell about how we develop cognitive mechanisms underlying intentional states. The way our cognitive

mechanisms are aimed at the environment is determined by this environment at least as much as it is determined by the make-up of the mechanism itself, or more specifically: the fact that the mechanism exists and has such-and-such a make-up depends on (= is explained by) the environment, while the fact that the mechanism operates in such-and-such a way (i.e. provides a certain output) given a specific environment (and a certain input), depends on its make-up. Cognitive mechanisms, as implemented in our nervous systems, are embedded in the environment in this way. Insofar as we can describe some of these mechanisms as being aimed at certain features, this property of being aimed at them is only explainable by invoking features of the environment. And underlying our intentional capacities is a pruning (using Buller's term) or a trimming (using Quine's term), which shapes our internal features as a consequence of interaction. That is: the fact that we have intentional psychological theories and that they are used in our communities enables us to have cognitive mechanisms which can operate on inputs distinct from those in communities in which this isn't the case, and they can operate on those in ways specific to our community. My favourite examples include being afraid of stock-market crashes (compare I.4.5 and I.7.5) – an intentional state depending on a wholely community-made phenomenon.

And returning to the toad's cognitive mechanism: Its being aimed at elongate apparations, and ultimately at worms, is explained by these elongate apparations usually coinciding with the presence of worms, and with worms being sources of nourishment. There are two ways of looking at this: Either the mechanism is described as being sensitive to elongate apparitions, which can be described in terms of intrinsic properties (i.e. sensory stimuli), or it is quasi-(hemi-semi-demi-)intentionally described as being aimed at worms. In the first case, the environment is invoked to explain why this mechanism has evolutionarily developed in this specific way in the first place. In the second case, the environment is invoked to interpret the mechanism's interaction with the environment as the toad's predatorial intent being the capture of worms. That it is aimed at worms is, as in the case of *Simon's Ant* (cf. Haugeland 1995: 209), not an intrinsic feature of the toad, but a result of its being placed into the environment which is interlocked with the specific cognitive mechanism at the toad's disposal.[54] Intentional state descriptions should be taken as analogous in the

[54] It should be noted that if cognitive features are enacted as they are in the example of Simon's Ant (i.e. if the complexity of the path an ant takes is due to the interaction of a simple algorithm embedded in the ant's brain with the location's geology, see footnote 53), the most promising way to entangle this interaction is by performing an investigation into the underlying neural properties. Since this is possible in the case of the ant (and the toad; see section II.6), we can arrive at the insight that there is in fact no additional encoding of the path (or the worm) in their brains. In humans, this disentanglement will be much harder, since the primary insight into the neurological correlates of cognitive function are acquired through neuroimaging. As Sullivan points out, "[i]dentifying the neural basis of a cognitive capacity is assumed to be achievable by correlating (a) subjects' behavioral performance on experimental tasks or their subjective reports with (b) measurable brain activity"

following way: Saying that someone has a specific intentional mental state is to say that they (1) have a suitable cognitive mechanism at their disposal, which may well be described in terms of intrinsic properties, such as properties of their nervous system, but also that (2) there is a psychological law which implies that the mechanism is interlocked *reasonably* with the environment, which is described as an extrinsic property. Without the psychological law, there would be no content.

Psychological laws depend on extrinsic properties because they rest, among other things, on conceptual relations between mental states (see the very beginning of section I.6.6). The content of such mental states, which is what determines the explanatory roles they can play, in turn rests on what Quine calls observation sentences, which express external circumstances. Observation sentences – i.e. sentences whose truth depends only on observable circumstances – are "the primitive source of the idioms of belief and other propositional attitudes. Without the aid of the observation sentences, it is not possible to make statements about the beliefs and values" (Bhat & Sahu 1998: 403). "Observation sentences even in this ultimate sense [as observables of a whole speech community] are reports not of sense data still, but of ordinary external circumstances (...). Many are nevertheless learned by direct conditioning to sensory stimulation, and all of them could be. Hence their epistemological significance as a link between our sensory stimulation and our theories about the world" (Quine 2008: 369).

However, this point does not imply that the content of mental experience stands in an inferential relation with non-conceptual states – a potential fallacy harshly criticised by McDowell as succumbing to the "Myth of the Given" (cf. McDowell 1994; for a concise summary see Zeglen 1991: 117-128). Quine carefully avoids this fallacy when pointing out that "[s]ome of my readers have wondered how expressions that are merely keyed to our neural intake, by conditioning or in less direct ways, could be said to convey evidence about the world. This is the wrong picture. We are not aware of our neural intake, nor do we deduce anything from it. What we have learned to do is to assert or assent to some observation sentences in reaction to certain ranges of neural intake. It is such sentences, then, thus elicited, that serve as experimental checkpoints for theories about the world. Negative checkpoints"

(Sullivan 2015a: 33). Imagine the only way to analyse the ant's neural properties were by neuroimaging: we would correlate the observable behaviour (i.e. the complex path) with its neural activation. Nothing gained from this correlation would straightforwardly tell us what the correlated activity actually represents: a simple "if path is obstructed, go right"-algorithm, or a complex path-description. Considerations of parsimony could lead us to suspect that if a simple algorithm could do the required work, then that is what is neurally embedded. However, this hypothesis is not a consequence of the imaging data, but of theoretical considerations. I believe this mirrors what is currently happening in the field of enacted cognition, where hypotheses about sparse encoding are guiding both empirical investigation in biological nervous systems as well as the construction of artificially intelligent systems.

(Quine 1993: 110 f.). The picture of how we have learned to assert or assent to specific observation sentences in reaction to certain ranges of neural intake is to be modelled after Davidson's triangulation (see section I.7.4), where giving content to one's and others' mental states depends on social and environmental interaction.

A concluding cautionary remark: While this interlocking of mental state and environment works somewhat analogously to the toad's mechanism being evolutionarily interlocked with its environment, I have suggested that it also goes beyond it (namely, by way of mutual interpretation in communities) and thus needs to be distinguished from it. To say that something is reasonable is, of course, different from claiming that it is evolutionarily fruitful. Claims about evolutionary advantages are not verified by saying that a specific mechanism serves an advantageous goal – this is but a heuristic assumption for formulating hypotheses about evolutionary origins –, but also that it in fact *was* this specific advantage which led to its persistence in a specific population of the mechanism's bearer's ancestors (compare Millikan's and Neander's quotes in my "derivability argument" above, as well as section II.7.1.). Claiming that something is rational is obviously very different: To say that an *action* is rational is to say that, given the agent's knowledge about the relevant instrumental relations between action and desire, the agent can expect the action to bring about the consequences she desires. Claiming that a *belief* is rational is to say that it is largely coherent with other beliefs held by the person in question, and that the belief stands in a certain relation to what we know about the agent's knowledge about the world (which we will often try to learn by observing how the agent interacts with her environment, compare section I.7).

Now, to say that someone has a specific mental state with its respective content is to assume that they are to some degree rational, but it would be absurd to demand that any content we can assign to her contributes to her evolutionary fitness (and is thus evolutionarily determined). The idea is rather that, insofar as the general capacity to have intentional states has such advantages, it can be explicated evolutionarily. For example, the connection between evolutionary favour and epistemic faculties such as the capacity to form true beliefs has been repeatedly made, be it by Quine ("Creatures inveterately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing their kind", Quine 1969: 126) or Dennett ("the capacity to believe would have no survival value unless it were a capacity to believe truths", Dennett 1971: 101). None of this is to say that the specific attribution of content to these intentional states is in any interesting way determined, or traceable back to, evolutionary constraints. I have cited the case of fearing stock-markets; I can go on to cite cases of desiring to be supreme ruler of Asgard, to draw unicorns, to live in the times of Isaac

Newton, believing to read a thesis, planning to go to Ireland, and so on. Some of these states may have roundabout connections to evolutionary constraints of our epistemic faculties; but their specific contents are probably as far removed from them as from the toad's endeavors to catch a worm.

I.9. Non-propositional Mental States

I.9.1. Non-intentional Psychological Explanation

Talking about "psychological explanation" is prone to cause misunderstandings, especially between philosophers and psychologists. This is because in philosophy, psychological explanation has been prominently explicated as intentional explanation, that is, the explanation of behaviour by invoking intentional mental states (see e.g. Cummins 1983 and Levine 1987). In psychology, this need not be the case. While going into the details regarding the nature and role of non-intentional states is beyond the scope of this book, there are some important connections to intentional states which I am going to explore briefly.

Bechtel & Wright (2009) give an overview over forms of psychological explanation which are "diverse and heterogenous" (ibid.: 113) but firmly at home in academic psychology: For instance, in psychophysics the Weber-Fechner-Law aims to give descriptions of regularities between physical and psychological phenomena with the "same formal rigor as description of laws between physical phenomena" (ibid.: 114). In information-processing psychology, tests have been devised for investigating whether subjects use serial or simultaneous search mechanisms when searching for items on a list (cf. ibid.: 115). In physiological psychology, it has been found that the stimulation of certain brain areas serves as a rewarding stimulus (cf. ibid.: 116). Also, damages to the ventromedial prefrontal cortex can be invoked to explain sociopathic tendencies (cf. Roskies 2003). Even though it may sometimes seem like intentional and non-intentional causes compete for what's *really* explaining a certain mental event, such results do not pose a challenge to intentional explanation. Acknowleding intentional causes as real causes does not mean precluding that "there is no difficulty in general in explaining mental events by appeal to neurophysiological or physical causes; this is central to the analysis of perception or memory, for example" (Davidson 2004: 180). And asserting that "there is no difficulty in general" does not mean that there is no methodological difficulty involved in coming up with specific explanations,

but rather that intentional and non-intentional explanation are readily reconcilable. Showing this is one of my aims in this section.

But first, let's look a bit further into what characterises non-intentional explanation. The explanatory relations between variables as just described are commonly referred to as psychological *effects* instead of laws (cf. Cummins 2000), and

> "appeals to effects are typically not explanatory. Instead, they serve to describe phenomena that in turn require explanation and elucidation – i.e., the explanandum. (…) The strategy described in many philosophical accounts is to explain empirical laws by deriving them from theoretical laws. (…) The challenge in applying this strategy to psychology is that [it is] unclear what the theoretical laws are to which one might appeal in explanations. An alternative is to appeal to the laws of more basic sciences (e.g., neurophysiology). Unfortunately, this approach is likewise problematic, as there are even fewer examples of relations called *laws* in physiology or biology. (…) [W]hen psychologists (as well as physiologists and many other investigators in the life sciences) offer explanations that go beyond the empirical laws or effects they identify, they frequently suggest that such explanations model a *mechanism* — i.e., a composite system whose activity is responsible for the target phenomenon. (...) [F]or the purposes of characterising information-processing mechanisms, the key point is that those mechanisms use (...) representations to coordinate the organism's behaviour with respect to or in light of the represented features of its environment" (Bechtel & Wright 2009: 118 f.).

Bechtel and Wright propose that such mechanistic notions serve to integrate "a variety of explanatory projects in psychology" (ibid.: 126), such as those marked by said diverse "effects", under one explanatory paradigm:

> "[M]echanistic explanations both explain lawlike regularities and appeal to other lawlike regularities to characterise the operations constituting the mechanistic activity (...) [cf. Glennan 1996]. [Their decomposition along these lines is] clearly reductionistic; (…) [however,] the organisation of components parts and operations, both spatially and temporarily, are crucial to a mechanism's activities, and this is not provided simply by lower-level laws or even knowledge of the component parts and operations themselves. (…) [W]hile mechanistic explanations are in part reductionistic, they also accommodate the emergence of higher levels of organisation and the need for autonomous inquiry into the regularities found amongst the denizens of these higher levels" (ibid.: 125).

As they also note, mechanistic explanation need not exhaust psychological explanation. Other explanatory accounts include dynamical systems theory and evolutionary theory (cf. ibid.; also see sections II.7.1 and II.7.2), and their use stems from connecting psychological and biological forms of explanation (cf. Braillard & Malaterre 2015; compare I.8.4 and II.2).

Such forms of explanation are clearly non-propositional, and, unlike common descriptions of non-propositional psychological phenomena such as feelings and experiences (see section I.1), the way they are described offers no hopes of connecting them to the way representations appear in the propositional form. So, even if some of the non-propositional forms of psychological explanation exploit *some* notion of representation, this notion won't be easy to connect to that of representation as used by intentional psychology. (The larger question in what way these non-propositional states or objects are representational/intentional will be dealt with in chapter II.) This gap between the two distinct forms of representation is underpinned by systematic differences between non-propositional forms of psychological explanation and intentional psychology: Namely, the psychological effects described non-propositionally, as well as the mechanisms underlying them, are usually independent of considerations of agential control, and thus, they are independent of the considerations of methodological restrictions of charity which is characteristic for the ascription of propositional attitudes (see section I.7.4). The way we experience the intensity of a stimulus, the way we search for items on a list, and the way our neural make-up reacts to stimuli is independent of agential considerations.[55] So, since agential descriptions have no systematic place in the kinds of non-propositional psychological explanation as just described, it makes sense that they are independent of ascriptions of semantic content, and in this sense also non-intentional.

However, non-intentional explanations can have a bearing on intentional/propositional states. For example, they can modulate behaviour which we commonly seek to explain intentionally, but whose modulation can thusly be explained non-intentionally. For example, the severity of moral judgments can be influenced by inducing emotions, particularly disgust.[56, 57] Intentional justification usually leaves room for a certain range of behaviour, so

[55] This insight is accommodated by "folk psychology", insofar as it does not justify holding people responsible for the way they are "hard-wired". Once we find out that a certain aspect of an action was not under agential control, holding the agent responsible for it is unjustified. While psychological results can push the boundary of what we hold an individual responsible for, they do not support a global skepticism regarding rationality and agential control. Rather, they point us to cases in which we err in ascribing agential control.

[56] See Wheatly, Haidt 2005, Valdesolo, de Steno 2006, Schnall et al. 2008a, Schnall et al 2008b, Jones, Fitness 2008, Horberg et al. 2009, Horberg et al. 2011, Eskine et al. 2011, Inbar et al. 2012.

[57] There is some controversy about whether the term "moral judgment" refers to a "cold" motivationally inert cognitive state or a motivationally charged state, i.e. whether one can judge X to be morally right without having any motivation to do X (under appropriate circumstances). *Motive Internalism* is the view that to make a moral judgment is to necessarily be motivated to act accordingly (see Roskies 2003, 2006 and Cholbi 2006a, 2006b). In

the fact that a moral judgment was influenced by emotions which in turn are not part of a rational justification basis need not outrightly conflict with the intentional explanation, but can complement it by specifying a value within this range. However, in cases in which we learn that induced behavioural effects exceed rational justification, agential ascriptions need to be retracted. (Except if the agent in question was wantonly negligent by ignoring that her judgment could be swayed by exposing herself to such conditions.) Generally, when tasks which should be under agential control are unduly influenced by a- or irrational mechanisms, explanation by non-intentional properties should not be viewed as undermining intentional explanation, but rather as specifically telling us what went wrong. Knowing that moral judgments are influenced by untidy workplaces supplies us with a good reason to see to it that relevant workplaces are tidy rather than with a reason to abandon all hope in morality or agential descriptions. Analogous lessons can be learned from studies on biases in general (although the question how much or whether anything can be done to counteract a bias depends on the details of the bias and the underlying mechanism; see footnote 39).

This complementarity of intentional and non-intentional explanation harkens back to the formers' normative aspect, which distinguishes it from other kinds of psychological explanation. If behaviour is found to be in conflict with norms of rationality, we are prompted to criticise the agent for not being in accordance with their reasons. Whereas if the behaviour is "in conflict" with other psychological effects that do *not* stand in a justification relation to actions then there are no grounds for criticising the agent. That is, there are no grounds for criticising someone for not conforming to interpersonally stable statistical effects which are unrelated to matters of justification and rationality. For example, if an agent does not conform to an effect such as the influence of induced emotions as mentioned above, the fact that she did not let herself be influenced rather appears as resilience to unwanted influence and is praiseworthy. The respective psychological research should not be viewed as contradicting intentional explanation, but as educating us about confounders of rational behaviour: "Time-pressure and emotional arousals (e.g. anger) are considered to be confounders for appropriate moral and rational considerations. (…) Research that shows that intuitions or emotions influence our moral judgments is more than welcome as it specifies the effect of confounders" (Triskiel 2016: 88 f.).

holding this view, one asserts that being motivated becomes part of the ascriptive basis of making a moral judgment.

I.9.2. Know-How

Many of us know how to ride a bike, how to operate a computer, how to get from our homes to our workplaces, but if pressed for a verbal description of these technical and motor skills, words might fail us. "You do it like this!" – pointing to an action, miming that action, training someone by making them repeat a certain action, that is what seems more proper in these cases. Of course, actions can be described, proper tool usage can be described, and experience may be simulated. Yet, even if words do not completely fail us, by coming up with these descriptions it often seems like we are only clumsily translating something which could be communicated much easier non-verbally. This *procedural knowledge*, this know-how, is often easily accessible mentally, yet in its original form it is not readily propositional (cf. Fridland 2015). Even if there might be a weaker notion of intentionality, which does not require propositional form, it is far from apparent how we could bend the concept around the notion of procedural knowledge. If I know how to ride a bike, does my knowledge refer to the bike? If I know how to operate a computer, does my knowledge refer to the computer? To how many things does your knowing how to operate a computer actually refer? In some way, if it refers to the skill, then why shouldn't it refer also to your hands, with which you possibly operate it? Does it refer in different ways to computers and your hands – does it explicitly refer to the former and only implicitly to the latter? Such questions stem from taking the concept of reference out of a context in which it is explicable and plunging it into rather murky waters. Reference can easily be explicated as symbolic reference – taking a signifier to mean something else –, but it is hard to see how our skills have anything to do with symbolic reference. While in some contexts, such as theatrical performances, we can symbolically operate a computer, our skill does not require symbolic reference in order to be a skill. The difference to thoughts and emotions, which always are to be about something to begin with, thus requiring symbolic reference, should be apparent.

I.9.3. Psychological Dispositions

Psychological dispositions such as being daring, quick-witted, mild-mannered, hot-tempered and the like are likely to be among the first things people will think of when asked to name some exemplary mental states, rather than propositional attitudes like beliefs or desires (or least those unspoiled by research in analytic philosophy of mind will). These

dispositions are the subject of differential psychology, and they are of particular interest to us as social animals because they explain variations in behaviour between different individuals. Within any given population, psychological differences will appear more striking than close similarities, which might explain why the associated mental states are quicker to be mentioned.

That is, the general psychological law that if people intend to attend a certain conference, then they are likely to show up there, might be exemplified by both Sam and Max, but Sam might be consistently more reliable in that regard than Max. Thus, to the people who know both Sam and Max, the fact that Sam is reliable, while Max is spurious, will be more striking to them than the law just mentioned – because there is, quite simply, usually no need to point out the obvious –, and the according differential psychological dispositions will enjoy more limelight, despite the respective intentional law being the basis for the evaluation of said dispositions. So, differences in dispositional ascriptions reflect differences in evidential bases for ascribing such states to different individuals. If the threshold for showing angriness behaviour is lower than average in a particular individual, we mark this threshold by saying that she's got a violent temper, but this means just the same as: typical conditions under which we are justified in anger do apply, but with differing intensity. Much the same goes for individuals whose psychological attitudes persist for shorter or longer times than average, and which we mark with adjectives such as spurious, resentful, or the like. These adjectives really describe properties of propositional attitudes or modifiers of their ascriptive basis.

While I do not wish to claim that all psychological dispositions are derived from propositional laws in the manner I just used as an example, I do suspect that in an overwhelming amount of cases, the connection will be rather intimate, and the worry that differential dispositions are either completely detached from the propositional form, or that they are more prominent or interesting per se, dominating the field of human psychology, should prove to be unfounded.

I.9.4. Qualia

One fundamental divide within mental phenomena is between those which are private experiential states and those which are not. Since the former are characterised by their qualitative aspect, they are traditionally called "qualia" (see Lewis 1929, Jackson 1982,

Dennett 1990 and 1991a). They are taken to be immediate experiences of mental states, as opposed to indirect experiences: we can experience another person's mental state by way of observation, or by way of whatever consequence it may have on us, but we do not experience it directly, immediately, introspectively, like we do our own mental states. If this notion of privacy and subjectivity is taken to be the primary characteristic of mental states, it results in a form of classic skepticism, characterised by conundrums such as: how do we know that another person *really* has a mind, as opposed to just *appearing* to have one? How do we know they aren't just a convincingly-made android, or some other form of soulless creature? Skeptic anxieties such as these suggest that we cannot rule out the possibility of *mental solipsism*, the view that there exists no other mind in the universe but one's own. It assumes that the only form of evidence for really knowing whether someone has a mind is its immediate qualitative introspective acquaintance. Since this acquaintance can only ever be given for oneself, mental solipsism cannot be ruled out. A skeptic view along these lines would, for mental ascriptions to be properly justified, not only require someone to show all outward signs of having a mind, but for them to also have the subjective experiential quality that comes with it (see section I.7.2).

In this section, I will show how a commitment to the reality of qualitative mental states can be reconciled with an antiskeptical position as laid out in I.7.4. (A quick reminder: in that section I showed that private states are not exclusively characteristic for our picture of mental states, and that mental states, at least insofar as they are intentional, are public states and intertwined with intersubjectivity and objectivity.) Like Shoemaker, I believe that "it is essential for a philosophical understanding of the mental that we appreciate that there *is* a first person perspective on it, a distinctive way mental states present themselves to the subjects whose states they are, and that an essential part of the philosophical task is to give an account of mind which makes intelligible the perspective mental subjects have on their own mental lives" (Shoemaker 1996: 157). What I am going to argue for is that this view does not conflict with the view that mental states are a public matter.

Crucially, taking a skeptic stance as outlined goes beyond merely stating that mental states are accompanied by an immediate experiential quality for those having them. Accompaniment alone could be interpreted in several ways that do not imply other-minds-skepticism: the relationship between qualia and mental states could be accidental, coincidental or functional, and each of these relations would not pose the skeptic's challenge. That is, it may be the case that the human mind just happens to be fashioned in such a way that immediate experiential qualities accompany mental states, but only insofar as this

accompaniment is an *accidental* fact about the world we live in, and we could well envision our minds as having been fashioned in a different way. Or, as a second alternative, it might be the case that immedate experiential qualities simply happen to temporally and/or locally *coincide* with mental states, but are otherwise completely unrelated to them – just as my neighbour might coincidentally take out her trash at the exact same times as I do, but without us ever interacting. Or, as a third alternative, we could take these experiential qualities to serve a *function*: If an agreeable feeling accompanies my belief that I have helped someone dear to me, then it is more likely for me to want to repeat helping others than if the accompanying feeling were to disagree with me (cf. Strohminger 2015). Similarly, the function of pain is to result in an immediate averse reaction to what caused the pain. This reaction may in many instances be controlled, sublimated or refrained from and in such instances remain unobserved; but the general function obtains. In this sense, experiential qualities may have properties which serve a specifiable function (this function could be social in nature, related to self-preservation, or what have you). Now, since functions can be implemented in various ways, any such function, in which an experiential quality plays a role, could also be fulfilled *without* experiential qualities playing any functional role (even if we were to find out that we all happen to be wired in a way that it is exactly those experiential qualities which are conducive to fulfilling our mental functions). To give another example, one prominent hypothesis about the nature of consciousness is that it is a functional property of the brain, making mental content available through a specific pattern of neural activation, thus allowing us to react more efficiently to our surroundings. But this functional "access consciousness", so described, could very well be implemented without requiring any qualitative states and thus, these would not appear in an explanatorily valuable description of the former: "To explain reportability, for instance, is just to explain how a system could perform the function of producing reports on internal states. To explain internal access, we need to explain how a system could be appropriately affected by its internal states and use information about them in directing later processes" (Chalmers 2010: 6; also cf. Block 1995: 229).

These views are all compatible with the assertion that there are immediate experiential qualities associated with our mental states, meaning that this assertion alone does not lead to skepticism by default. That is, if we construe qualitative mental experiences as accompanying certain mental states, but don't construe qualia as the be-all and end-all of the mind, then we do not have to accept skeptic criteria for what qualifies as having a mind in the first place. We can very well be realists about qualia without having to seriously consider mental solipsism.

In effect, skeptic views require us to believe that mental states are *only* mental in virtue of their experiential quality, and we have to concede no such thing. I take it as evident that we do have qualitative states, because in fact there *is* a taste that goes with my eating chocolate, there *is* a feeling that goes with wind on my skin, and there *is* an elation that goes with my listening to Beethoven. However, it is just as evident that mental states can be ascribed without having any direct experiential access to the qualia that accompany them – I know for a fact that many people share my feeling of elation, or my taste, without needing direct introspective access to *their* qualia, and that mice generally fear snakes, without having to concede that the mental experience of a mouse is similar to mine. That is, even those mental states which are subjectively characterised by a certain qualitative experience are also characterised by intersubjectively accessible features. (I dare anyone to come up with a mental state which is so radically private that it cannot even in principle potentially be connected to publicly observable features; and I double dare them to take this mental state as characteristic for the human mind in general.)

It should also be noted that many mental states we have immediate access to are not even accompanied by any characteristic qualitative experience, and no qualia whatsoever are required for identifying them as the mental states they are. For example, I have immediate access to my belief that 2 + 2 = 4, or my belief that Socrates was male – yet it would be hard for me to associate these beliefs with any characteristic qualitative feeling. Thus, what is mental does not generally coincide with what has a specific experiential quality.

The crucial follow-up question is whether the privacy and subjectivity of qualia render them opaque to scientific research – and if not completely, to which degree science can investigate them, or whether we need to devise some ingenious new research strategy. David Chalmers appropriately dubbed questions having to do with these qualitative aspects "the hard problem of consciousness", since, unlike in the case of "easy questions", at present we do not even have a vague notion of how such problems are to be solved (cf. Chalmers 2010; see also section II.5). While in many areas of science we pursue a strategy of research leading up to an epistemic goal, it seems in the domain of qualitative mental states we lack even a general strategy. Thomas Nagel took the qualia's inherent subjectivity as keeping them from currently being the subject of objective research (cf. Nagel 1974). However, I do not think that the distinction between qualitative and and non-qualitative aspects of mental states coincides with the distinction between what is beyond and what is within the domain of proper science. Rather, I would like to offer what I hope are some helpful distinctions *within* the domain of qualitative states, which offers room for their exploration.

The qualitative experience of mental states can be explored in three different ways: By way of phenomenological description, by way of their intentional object, and by way of a correlation with objective facts. While we have to accept that the experienced quality itself cannot be intersubjectively shared or accessed, this does not have to keep us from scientifically investigating the objective features that come with it. While it is true that some qualitative experiences are rather unique and impossible to relate, others are common and have intersubjectively evaluable properties. The former are usually associated with exclamations like "you simply have to be there!", "you just have to taste it!", "you won't believe what it sounded like!", and so on, implying a lack of helpful description, thus eluding our objective grasp (although not completely, since the sensory modalities in question are always being explicitly referred to, and the assertion that the content is in some way "indescribable" does communicate something quite specific). The latter can be traced to assertions like "the rides at the Oktoberfest don't hold a candle to those at Coney Island", "grapes are sweeter than lemons", "Hitchcock's movies are more emotionally immersive than Kubrick's" and so on. The fact that assertions like these are alive and well means that we readily acknowledge that certain kinds of subjective experiences are shared by many (if not all) people, and that they can, at least to some degree, be compared both inter- and intrasubjectively. And the degree to which they are comparable isn't even supposed to systematically differ from the comparability of external, objective, quantifiable facts. If I take a trip today, and one next week, I'll probably be able to pick the one I liked better, and if I talk to someone who took the same trip as the one I took, we are likely able to exchange our opinions and give reasons as to why we thought it was a nice one or not. Judgments of personal taste, especially when they're not marked as such, may occasionally dilute the intended objectivity, but they do not render the comparison itself a hopeless endeavor. To give a further example, judging whether the screen in a movie theater is dimmer during a 3D presentation is a common and well-justified subject for objective debate among passionate movie-goers, all the while resting on subjective experience. (Also, qualitative experience can not only be tied to intersubjective concepts qualitatively, but also quantitatively, such as in psychophysics, see I.9.1.) The list of potential examples for productive talk about subjective experiences goes on and on.

Delving into the semantic features underlying this talk about qualia, it should be noted that any terms referring to them could not refer to radically private phenomenological states if their privacy implies their singularity (i.e. that it is conceivable that only one person ever has them) or their being completely detached from public aspects or correlating conditions which

we could exploit to make them intersubjectively accessible. We could not understand words which are used strictly idiosyncratically, and the practice of conversation would break down if, where observable behaviour is guided by questions of semantics, semantic rules would wildly differ or change arbitrarily between speaker and interpreter, and where they could not be fixed in relation to intersubjectively available conditions.

Rather, a considerably more plausible view is that in expressions referring to qualitative mental states we should expect to find phenomenological and public aspects intertwined. In the following excerpt from his famous "private language argument", Wittgenstein asks:

> "If I say of myself that it is only from my own case that I know what the word "pain" means – must I not say the same of other people too? And how can I generalise the *one* case so irresponsibly?
>
> Now someone tells me that *he* knows what pain is only from his own case! – Suppose everyone had a box with something in it: we call it a "beetle". No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle. – Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. – But suppose the word "beetle" had a use in these people's language? – If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a *something:* for the box might even be empty. – No, one can 'divide through' by the thing in the box; it cancels out, whatever it is.
>
> That is to say: if we construe the grammar of the expression of sensation on the model of 'object and designation' the object drops out of consideration as irrelevant" (Wittgenstein: PI §293).

Here, Wittgenstein criticises the Cartesian model (see I.7.2) as misguided. If the meaning of a concept that is used in public discourse is radically private ("No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle"), then the meaning is bound to "drop out of consideration". "What this shows is that it cannot be correct to construe the "beetle" language game on the model of 'object and designation'. On this model, the object is crucial to the use of the designating expression; it makes a difference to the use. So where the putative object makes no difference to the use of a term, it makes no sense to insist that the grammar of the term is that of a designator" (Williams 1999: 32). The initial persuasive power inherent to Cartesian skepticism turns out to rest on a confusion of treating supposedly radically private things as contributing to public discourse. For there to be

the word "pain", we must already assume that it was learned under public conditions, which implies that criteria of its use are based on observable evidence. So, at the very least, the "private" aspects of pain are intertwined with its publicly observable aspects. Again, this does not imply that there is no such thing as a private, subjective *feeling* of pain, or that there aren't any potentially unobservable "internal" aspects to pain; just that we cannot neglect its "accoutrements" (ibid.: 30 ff.), that is, the external, public aspects of pain which pain-discourse and pain-ascriptions are necessarily based on.

Now, the one question left to answer is: How do we fashion intersubjectively available concepts which have subjective states as their referents? While it is true that only one person ever immediately experiences their own qualitative subjective state, there are forms of intersubjective access. Indirect as they may be, they are sufficient for intersubjective conceptualisation. Subjective experiences are tied to behaviour after all, and having them stands in a systematic relation to observable behaviour. And when it does not, self-reports can bridge the gap between ego and alter ego. For example, imagine Tyler going down the street with Marcia, when he sees someone spitting on the street, narrowly missing his shoe – something which Marcia remains unaware of. Tyler consequently experiences a bout of cold anger, causing him to exert some grip on himself in order not to let Marcia be on the receiving end of his anger. So, even though there may be no observable behavioural difference to Marcia, Tyler may, through his acquaintance with Marcia, very well be justified in expressing his anger to her, telling her that he needed to exert some degree of self-control. Marcia may trust his self-report, knowing that Tyler is a trustworthy person, and she knows that he is because he often behaves in accordance with what he says. For instance, whenever he lets her understand that he is angry, she may observe him to react more tensely in stressful situations. Just as in our present case: If Marcia does not know right away whether Tyler is being honest or joking, she may continue to observe his behaviour more closely over the next few minutes. It is through this web of various forms of intersubjective access that we ultimately come to the result of unwaveringly believing that most human beings, and even many animals, share some subjective qualitative mental states, even though there can never be more than one person having their own experiences. And through this interplay of different forms of access, this belief has much stronger support than all the nagging doubt that has been proliferated by skeptics and solipsistic thought experiments through the ages, by stories about zombies, clones without souls, human-like robots, and the like.

Thus, the initially puzzling question of how we manage to arrive at objective facts when starting out from subjective experience can be answered three-fold. Firstly,

phenomenological description is a form of description after all, and as such, it requires language. Language is, among other things, a way of connecting subjectively perceived stimuli to intersubjectively accessible and employable entities: concepts. As far as we are able to describe our experience using established concepts, we can rely on intersubjectively stable dispositions to react similarly under similar conditions (cf. Quine 1960: 7), ultimately arriving at a discourse allowing for objective topics. Our talk about subjective experience qualifies as one such form of discourse, since our concepts about sensory perception, impressions and tastes have proven enormously stable: we talk about our experiences all the time, and even though we know that we may not be able to directly share them, we can justifiably expect our talk about them to be understood (and we can also to some degree expect to evoke similar sensations as those we have experienced ourselves in others, by bringing someone whom we know to have tastes and dispositions similar to our own into a context which shares relevant conditions with the one which evoked our own sensation). We can expect others to understand what we mean when we say that we feel elated, indignated, or that we enjoy the taste of chocolate. So, wherever there's a concept, there's something intersubjectively learnable.

Secondly, insofar as we can express our experiences as having an intentional object, we can relate them this way: My experience of tasting chocolate is comparable to someone else's experience of tasting chocolate in virtue of their being tasting sensations *of the same intentional object*. These intentional objects are usually individuated externally, and then there should be no doubt that they are intersubjectively accessible. Some experiences may be purely internal – they may be experiences of stomach ache, or even of some internal state which cannot be related to any external cause or condition. In these cases, intentional objects are only helpful insofar as they can yield generalisable concepts, such as the concept of stomach ache. But when they are external, which should be the case in an overwhelming amount of subjective experiences, then the following holds: As much as our experiences when riding the ferris wheel can differ, they are both experiences *of riding the ferris wheel*. Any person can in principle make them, and any person can in principle repeat them. Thus, the intentional object is a part of analysing the subjective experience: It is fundamental to the experience of riding the ferris wheel that it was an experience of riding the ferris wheel instead of, say, tasting chocolate. And even qualitative differences in making these experiences are in principle accessible: If my riding the ferris wheel made me feel good, and someone else bad, then feeling good and feeling bad can be explicated in terms of generalisable concepts (that is, they are applicable over different persons and situations – feeling good or bad can apply to any number of situations and be experienced by many). Thus, whenever some parts of expressions

relating to subjective experience cannot be analysed in terms of external intentional objects, they may still be analysed in terms of concepts referring to generalisable properties.

And thirdly, the most common way of referring to qualitative experiences is to relate them to objective facts, be it by way of causation or correlation: We talk about the sensation we have when seeing the color red, when tasting chocolate, when feeling warm. Whether something is red, whether it is chocolate or whether it is warm is closely tied to objective facts. There are, of course, cases in which this connection seems more loose or more controversial than in other cases: for chocolate will always remain chocolate (marginal cases of chocolate/vanilla hybrids or similar ambiguities excluded), whereas whether it is warm is relative – i.e. a more or less arbitrarily chosen point or area on an objective scale of measurement –, and whether something is red can be up to debate: we can point to a certain objective property of light which under some optimal condition causes most of us to see a color we call red, but whether that objective property is to be identified with the experience of seeing red is itself not an objective fact. At best, their identity is guaranteed by our pragmatic decision that it should be identified in the way as is usually being done. On the other hand, whether we're all "really" seeing red when being exposed to a certain wavelength is usually not up for debate (i.e. save for some pathological cases), but only whether we all use the concept "red" correctly. What is required here is nothing more than to consistently use the concept whenever we are prompted by the according wavelength stimulus; the only way in which subjectivity sneaks in is by the subjective consistency of the stimulus, i.e. by judging whether the wavelength elicits the same stimulus now as in previous instances. On the one hand, comparing my belief today that 2 + 2 = 4 to my belief yesterday that 2 + 2 = 4 does not at all depend on subjective sameness of experiential quality, since it constitutes a perfect example of an intentional mental state whose content does not refer to any phenomenal quality at all. On the other, objectively comparing my sensation of redness today to any sensation of redness I might have had in the past seems completely hopeless for lack of any objective standard. In the absence of such a standard, the subjective comparison which we are left with seems like a leap of faith, pragmatically validated by external reinforcement in every case in which our comparison is deemed successful by our peers, who in turn have taken the same leap of faith in the accuracy of their senses. Said comparison itself is well beyond objectivity, and there can be no epistemic grounding for it; it can be causally explained by our evolutionary heritage, insofar as we need to have a built-in way of judging sameness of stimulus (see Bhat & Sahu 1998: 406), but that is really not what is at stake here; because once this subjective comparison has passed its intersubjective test of showing consistent

prowess at using the concept "red", our sensations of redness are objective insofar as they are sensations *of redness* – i.e. of something which can be referred to by a concept which evidently works intersubjectively. We often have reason to believe that people are actually seeing red when they're saying they do, even in the absence of independent proof (e.g. on the phone), since checking their reactions to external conditions is the basis for ascribing the mastery of such concepts to them (compare the concept of triangulation in I.7.4).

Neither is the question whether qualia are "really" intersubjectively alike, comparable or similar of pressing importance. Expressions which can be directly bound to non-linguistic stimuli, much as the uttering of "lo, there is a rabbit" can be bound to the presentation of a rabbit-stimulus, thusly have what Quine calls a "stimulus meaning" (Quine 1960: 32-36, see also section I.7.4). Quine has argued for the irrelevance of likeness of stimulus meaning for observation sentences, and I believe that if in the following quote you substitute "stimulus meaning" for "experiential quality" and "observation sentence" for "concept" (i.e. intersubjectively available semantic content), the argument works much alike:

> "The view that I have come to, regarding intersubjective likeness of stimulation, is rather that we can simply do without it. The observation sentence 'Rabbit' has its stimulus meaning for the linguist, and the observation sentence 'Gavagai' has its stimulus meaning for the informant [i.e. the speaker whose utterance is to be translated into the interpreter's language]. The linguist observes natives assenting to 'Gavagai' when he, in their position, would have assented to 'Rabbit'. So he tries assigning his stimulus meaning of 'Rabbit' to 'Gavagai' and bandying 'Gavagai' on subsequent occasions for his informant's approval. Encouraged, he tentatively adopts 'Rabbit' as translation" (Quine 2008: 371).[58]

To sum up, I take it that inner experiential states, insofar as they matter intersubjectively, are individuated intentionally; that is, they are usually individuated by their intersubjectively accessible content, which can be expressed propositionally, such as "the feeling that it is warm", "the sensation that I am perceiving the world as a bat would", etc. That is, insofar as subjective states can be individuated by way of propositional content, they, too, have a firm stand both in the realms of the intersubjective and the objective, even without ever being "felt" (i.e. directly accessible) by more than one person.

[58] One could object that Quine's argument deals with stimuli which are intersubjectively accessible to begin with. However, the point is that the private "meaning" only (at best) indirectly determines the intersubjective meaning. Rather, it is the intersubjectively accessible stimulus (which is related to the qualitative experience) which forms the basis of communication. In this sense, we can do without intersubjective likeness of stimulation – i.e. without asking whether the stimulus, as it is subjectively perceived, is "really" alike.

There is still the nagging doubt whether some qualia could actually be individuated by their quality alone, but I have some difficulty in coming up with examples for these cases, possible expressions likely being "the feeling I have now" or "this vague fuzzy sensation I sometimes get", or the like.[59] But these expressions do have at least *some* external, intentionally expressible correlates, such as a time index (e.g. "now") or another external qualifier ("fuzzy"), allowing for indexical or demonstrative reference. Some of these may be but presymbolic or -linguistic, in the sense that they only lack a proper expression, but are not inexpressible per se. Some others may be intentional (insofar as they have indexical or demonstrative content), but genuinely non-linguistic, non-propositional and non-conceptual. I will come back to this open question in section I.9.5.

None of what I have said in this chapter was designed to meet the skeptic challenge of fundamentally doubting that any other mind except one's own really exists. Rather, I have suggested that the view from which skeptic doubts follow does not lend itself to begin with. What I have said is that qualia come with a host of objectively evaluable, intersubjectively available properties, and that we can refer to these properties symbolically – by making them the content of symbolic representation. To briefly venture into science-fiction territory: We should expect a computer with suitable algorithms and sensory and/or motor modalities to be able to perfectly learn our language, down to expressions referring to qualitative experiences, even if this computer can have none. Thus, speaking our language cannot constitute proof that one has these experiences; but we should not expect any language to be able to refer to something inherently private, as such a language could not be learned.[60] Yet all of us constantly use language to express our experiences, and we can do so because they are linked to intersubjectively available facts. We know this to be true, yet there is no compelling reason to believe that we should take them to be all there is to mental states. Our inner experiences are linked to intentional objects, to concepts and external things because they have a purpose; because none of our inner feelings is fully behaviourally inert. This is not to follow the behaviourist dogma that all there is to human psychology is behaviour, but rather to acknowledge that the cognitive processes underlying our behaviour have a purpose which goes beyond their own private experiential realm. Our mind is not self-sufficient. Rather, it exists to connect us to the world.

[59] Compare Quine's saying that "just a few [observables from scientists], such as the indescribable smell of some uncommon gas (...) would resist reduction" to observables of the whole speech community (Quine 2008: 369).
[60] Compare Wittgenstein's "an 'inner process' stands in need of outward criteria" (PI §580) – meaning a private mental state cannot be adequately discussed without public criteria for identifying it. If there is something purely private about mental states, then they are irrelevant for the meaning of mental terms, since meaning is tied to public linguistic behavior. Mental states we can talk about cannot be strictly *private*. And vice versa: if we consider something to be strictly private, it follows that we cannot talk about it (compare section I.7.2.).

I.9.5. Non-conceptual Content

There has been a controversial debate about whether non-conceptual content actually exists and what it could exactly amount to, and not just in the domain of qualitative experience. For example, "[i]t is compelling to think of (…) [some] beings as having experience (...) [who] are unable to communicate thoughts to us; we are unable to understand – from the inside – how they are responding to the world; we are unable to impose our world on them" (Cussins 1990: 134). It is tempting to take the intentional stance toward beings which "*need not have* those concepts" (ibid.) when these ascriptions yield some explanatory surplus for us (cf. Dennett 2007: 87 f.), such as in cases of ascribing mental states to robots, thermostats, "very young human infants (before the acquisition of the object *concept*, say), or very senile people, or certain other animals" (Cussins 1990: 134). On the other end of the spectrum, Davidsonian holism commits those who adopt it to denying the ascription of content to beings who do not possess concepts, since saying that a dog believes a cat to have climbed a tree would require attributing to the dog "many general beliefs about trees: that they are growing things, that they have leaves or needles, that they burn" (Davidson 2001b: 98) – which, of course, is more than doubtful.

There is no harm in admitting that ascribing an attenuated form of belief to a thermostat is justified by its doing some explanatory work (within the boundaries of the attenuated ascription, see section I.7.5), even while admitting that the thermostat itself possesses no concepts at all. Here, saying that the thermostat quasi-believes that the room is too cold does not imply that the thermostat possesses the concepts "cold" or "room". This is because the ascription is not made "from the inside" (as in: what the thermostat is supposed to be thinking), but just to highlight the connection between the external world and the thermostat's reaction (namely, heating the room).

But are there mental states which we cannot even in principle ascribe conceptually or propositionally? Cussins invokes the example of a perceived sound: "Evidently the content is indexical or demonstrative since, were we to express the content in words, we would say that perception presents the sound as coming from "that location," or "from over there"" (Cussins 1990: 143). He judges this form of content to be non-conceptual on the grounds of its being indexical or demonstrative, not of its being a qualitative experience (ibid.: 139 f.). Now, Cussins may be correct depending on the framing of his example, but generally, if we were after a proper description of perceived acoustic content, I believe it would be much more plausible to describe it in terms of its qualities rather than its source. Of course, this depends

on how apt we are at expressing the quality of our perceptions – but the same thing also applies to expressing our beliefs: if we are not apt at using, say, concepts from mathematics, we might refer to a certain mathematical proof in an indexical or demonstrative way ("the proof Professor Ein was lecturing us about in Zurich"). But this does not keep the proof from being expressible in a proper conceptual way (such as "the proof was about the Fibonacci sequence").

Analogously, acoustic engineers can be apt at describing qualities of the sounds they hear, and so the resulting description would very likely neither be indexical nor demonstrative. For example, envision someone who can describe their acoustic experience in terms of notes, harmonies or acoustic frequency: either of these would qualify as concepts if we think of a concept as something that "divides up the world into objects [or] properties" (ibid.: 134), "which presents the world to a subject *as* the objective human world about which one can form true or false judgments" (ibid.: 133) and which is formed "relative to a theory" (ibid.: 134). Notes, harmonies and frequencies can also figure in propositional belief, desire and intention ascription just as any linguistic concept ("I desire to hear her sing in A minor", "I believe I heard a sparrow's trill"). So we can grant that if some content can only be expressed indexically or demonstratively, it cannot be conceptual; but, at least from Cussins' example, we cannot conclude that there *is* any content that is only expressible indexically or demonstratively. I have failed to come up with such examples, but of course I cannot rule out that they do exist. If they do, they are well beyond the scope of my claims in this book.

I should stress that my own view on this topic is limited by what I think can reasonably be called "content", and of course there are alternative views available (for a comprehensive overview, see Gunther 2003). As evidenced in the previous sections, my view is bound to matters of symbolic representation rather than a complete inventory of cognitive operations in humans. People who believe that mental content is the proper form of content and that non-mental content is derivative (see section I.4.4) are likely to have a different conception of content: For them, something's being mental or cognitive is by itself already a good heuristic for its having content, while I believe that something's being representational (or at least standing in a necessary relation to symbolic reference) is the proper criterion for its having content, insofar as such phenomena are the only ones which present us with the more tricky aspects of content-ascriptions (see ibid.). I have previously mentioned Searle as advocating the former view ("All linguistic meaning is derived intentionality", Searle 2000: 93), but in fact, Cussins does so too: "There are derivative uses of the notion in application to the communicative products of cognition, such as speech, writing, and other sign-systems (…)

but these uses must ultimately be explained in terms of a theory of the primary application of content in cognitive experience" (Cussins 2003: 133). However, as I have argued in section I.4.4, there is no fruitful way of characterising some of our interactions with the world as intentional if not by invoking symbolic practice. The catch is that without it, all interactions between cognitive apparatus and external world would boil down to causal processes. But since we do have symbolic representation at our disposal, we can sort these causal processes into two different processes: those which are intentional, and those which are not. If mental intentionality were primary and symbolic intentionality derivative, we should be able to sort our cognitive processes into intentional and non-intentional processes independently of symbolic properties; but of course we can't. The fact that we have mental content implies that we have a theory of the causal laws underlying it, and that these are individuated by reference to objects in the world. And there is no way of formulating theories without exploiting symbolic properties.

If all there was to our fearing snakes was our cognitive way of dealing fearfully with snakes – of processing stimuli associated with snake-appearances, of outputting aversive behaviour, of forming traumatic memories of snake-incidences etc. – then these wouldn't be intentional. Why? Because they would be completely describable in non-representational causal terms – down to said memories –, and there would be no need at all to come up with intentional terminology. They only become intentional once we come up with something that signifies a snake, and once we come up with the cognitive apparatus enabling us to deal with these signifiers. Whether these be grunts, words or signs – they need to signify snakes. Once they do, our mental life is enriched by intentionality. Until then, our "internal representations" merely stand in a non-representational causal relationship to what they purportedly are about (and while it is sometimes said that this is in effect all it takes, I beg to differ – see II.4 and II.6). Thus, once again, mental intentionality is not primary to symbolic practice, even though symbolic practice depends on cognition.

So, I believe our best reason for speaking of a cognitive state as having content is that it is describable in intentional terms, and this requires conceptual description. There may be a wide variety of such descriptions available, since what our mental states can be about can be grasped using any concept from any theory. I have invoked examples from natural language, mathematics and music, but of course there are many, many more. So, finding that some property is non-conceptual is likely to point towards its not expressing any content in the first place. For example, in Cussins' case, the property "of having an active hypothalamus (...) is characterised by means of the concept *hypothalamus*, but an organism may satisfy the

property without possessing this concept. Therefore (…) [it] is a nonconceptual property" (ibid.: 135). And having a hypothalamus is "not a content property, obviously" (ibid.: 160, FN 8). I already pointed out that I do not take the fact that there are beings whom we ascribe content to, but who themselves lack concepts, as implying that there is non-conceptual content. Rather, ascribing beliefs to infants, thermostats and robots only requires the *ascriber* to possess these concepts, and to employ these to explain the infant's, thermostat's or robot's reaction as an effect of some psychologically salient cause. (Since having the psychological concepts implies having the psychological theory, being able to apply the concepts and being able to explain the phenomena is much the same.)

The reasons why the psychological laws hold are different in these cases: In the case of the thermostat or the robot, an engineer has fashioned the system with the express purpose *that* the law should hold – so we ultimately explain the fact that the law holds via reference to someone who does grasp the relevant concepts. That is, the engineer's grasping the concepts is the cause for the thermostat's having the "belief" - and in this sense, it is much less attenuated than we might previously have suspected, when we thought that the ascription referred exclusively to an internal state of the thermostat. As I have insinuated in section I.6.4, to be able to explain something by psychological law is to explain it not just as an instantiation of an intentional law, but to also (at least implicitly) specify the cause of *why* the law holds (also see I.8.5). In the case of the robot and the thermostat, while the first condition is satisfied in virtue of the relation between context (i.e. the cold room) and the system whose behaviour we are to explain pseudo-intentionally (i.e. the thermostat), the second is satisfied only via reference to the engineer. In the case of the infant, both are satisfied by reference to the infant. Thus, the example of the infant catches our notion of "attenuated belief" much better than the thermostat, since in the infant's case, there is no full-fledged grasping of concepts involved in either explanation.[61] If a psychological law applies to an infant who has not mastered one of the concepts we use in the ascription that requires this law to hold, then the fact that the law holds cannot be due to the causal implications of having mastered the full-fledged concept (see I.7.5). So we say that we actually use *attenuated* concepts in our ascriptions (cf. Dennett 2007: 87 f.) in order to stress that the lawlike explanation implied by the theory holds without actually requiring the subject of the explanation to have any intentional/symbolic cognitive capacities related to the holding of intentional laws, and also to

[61] Although the cases might also work analogously if the infant's cognition depended crucially on having learned something; then we would have to ultimately explain its psychological property via reference to the external cause of his cognition, which might very well involve mastery of concepts. (This, of course, is the same difference as between explaining something as innate and explaining it under the social learning paradigm, cf. Levy 2004.)

highlight that adjacent psychological laws which might hold in the case of a person who has fully mastered the respective concept might not in this case.

I.10. Summary

Intentional psychology explains agential phenomena by ascribing to agents states which have content. With some exceptions, these states are classically taken to be so-called *propositional attitudes*: Attitudes such as beliefs, desires, intentions, etc. (sometimes also called "intentional modes") toward a propositionally formulated semantic content (I.1). Intentionality is the property of having such content, i.e. of referring to the objects, events, processes (etc.) or of being "aimed" at such objects, events, etc. (I.2). Content is individuated externally, in reference to matters beyond the agent, or, to be more precise: descriptions of intentional content take both internal cognitive properties of an agent into account as well as a relationship between the agent and the external intentional object (I.3). Having some kinds of intentional content and ascribing it to others both require social conventions which serve to establish what kinds of things refer to what kind of content, or, in other words, to establish symbolic representation (I.4.2). Causal chains between tokened symbols and their instantiated meaning can explain how the material parts of these symbols are associated with their specific meanings and how individual agents can acquire knowledge about symbolic representation. Yet, this explanatory relation between meaning and causality does not establish that representations are reducible to causes or effects (I.4.3).

While some believe that symbolic forms of representation are derived from mental intentionality (a view called "mentalism"), I argue that both of these are in fact interlocked, at least in those cases which have been typically invoked to argue for derived intentionality (I.4.4). In such cases, mental intentions can in fact only be properly explicated as mental states referring to symbolic meaning. Thus, they cannot be primary to matters of symbolic meaning. The hypothesis that thought is itself linguistic cannot make any headway toward clearing up matters of mental meaning, since it only pertains to formal ("syntactical") conditions for acquiring meaning. Further, both computationalist as well as connectionist principles can satisfy such requirements for neural processes to be interpretable as processing certain symbolic forms of content (such as linguistic content).

Since there are mental states which have content but on whom norms governing symbolic content have no bearing, I call such content "sparse", distinguishing it from "rich"

content which is had by those mental states on whom such norms do exert formative influence (I.4.5). Typically, rich content – which intentional psychology explanatorily invokes – is the kind of content which gives naturalistically inclined philosophers a run for their money, because it depends on non-natural properties such as rationality and normativity.

Mental states are theoretical objects insofar as they primarily depend on the explanatory roles they fulfill in psychological theories. Intentional psychological terms are also not defined over directly observable objects, but rather criterially inferred from (potentially) observable phenomena. Consequently, I suggest that debates about mental ontology should be led in terms of what is explanatorily valuable (I.5 and I.6.1). So, committing to the view that mental states are essentially theory-dependent does not mean committing to an antirealist view. If we focus on explanatory value, we also needn't be bothered by the distinction between "natural" kind-terms in scientific laws and those which are "unnatural". Of course, the distinction between kind-terms which denote things that depend on agents (such as mental states) and those which don't (such as chemical properties) is worth salvaging, but the insinuation that what is natural is in some way more real or scientifically more reputable is left by the wayside.

Intentional explanation is a form of lawlike and causal explanation: It invokes laws (i.e. general relations between projectible kinds) to explain singular instantiations of such kinds, and it treats positive instantiations as evidence supporting the respective general law (I.6.2.). Intentional explanation can be construed along the lines of the classical deductive-nomological model of scientific explanation: as a practical syllogism stating a general law and the instantiation of its antecedent as premises and the logically derivable consequent as the explanandum (i.e. post factum) or prediction (i.e. beforehand) (I.6.3). Beliefs, desires, intentions and similar propositional attitudes are statable in this syllogistic form, so that actions follow logically. Thusly, given an instantiation of the mental states invoked in the premises, the derived action, motivation, or reason to act, is explained or predicted relative to an agent. Such explanations do not merely depend on said states being relatable logically, but on being ascribable to the agent in question. To this end, the agent has to fulfill some minimal requirements of rationality: namely, her actions must be so systematic in relation to obtaining external circumstances as to be interpretable as being caused by veridical and consistent mental states (I.6.4). Also, in order for psychological laws to be explanatory, they need to be sufficiently generalizable (I.6.5) and relatively strict (I.6.6).

That mental states are theoretical kinds means that they are the kinds of things we can have theories about. Thus, they are systematically tied to observable phenomena (I.7.2),

especially those they explain: behavioural or behaviourally relevant phenomena. Yet, they are not restatable in purely behavioural terms (I.7.3). In building on the work of W.V.O. Quine, Donald Davidson argued that mental states are intimately tied to meaning, insofar as the acquisition of the means for symbolic representation depend on the ascription of such states, thus requiring the employment of psychological theories, and vice versa. He also ironed out how the rationality of an agent is necessary for her interpretation. I am following him in both regards (I.7.4). However, I do not follow him in tying having mental states to linguistic competence. I believe that ascribing mental states to animals, robots and thermostats can be well justified, even though we should mark such ascriptions as attenuations (I.7.5).

Building on this view, I aim to cement the notion that the individuation of an agent's mental content relies on more than her inner workings (I.8.1, I.8.2). While I follow both Putnam and Burge in their arguments for "broad" (i.e. externally individuated) content in special cases (I.8.3), I aim to cover more ground by providing two arguments for the claim that all kinds of rich content are individuated externally. The first says that even if all that is to know about the content of an agent's mental representations was determined by her intrinsic properties, we could not find out what these are if we did not look to matters beyond the agent (I.8.4). While this is an epistemic argument, its scope expands when considering that content can only come into play once we have developed a psychological theory. If no theory about content can assign it without taking matters external to an agent into account, then there is no "narrow" content (i.e. content which is internally individuated).

To make my second argument (I.8.5), I point out that an agent's inner workings always underdetermine the actual content of her mental representations, and that we cannot simply retreat to the view that her content should then be described as the specific *form* of underdetermination. This is because our mental states, no matter whether rich or sparse, do not merely refer to what is describable intrinsically, namely the proximal aims of cognitive mechanisms. For example, a toad's cognitive worm-detector is proximally aimed at certain elongate shapes and their movement. But detecting and processing such shapes and movements only makes sense when assuming that they are reliable indicators of nourishing external objects, and it is these external objects which explain the whys and hows of the cognitive mechanism. Sure, there *can* be intrinsic descriptions of such mechanisms, but these do not explain what theories invoking content are meant to explain. In fact, intrinsic properties can be explained without referring to content at all (and if we follow Fodor in holding that mental content ascriptions pick out intrinsic causal powers of agents, content as a kind-term itself vanishes). Rather, what intentional theories also explain – beside placing

functional constraints on internal causal processes – is a mechanism's presence and endurance, and they do so by pointing out or at least implying which functional aim (sparse) or norm (rich) has been instrumental for either or both. So, if we want an explanatorily valuable kind of psychological theory, we cannot stop at intrinsic analyses.

To conclude the chapter, I also took a cursory look at other prevalent kinds of mental states which are either representational but lack standard forms of being assigned content, or which are taken to be mental but possibly not representational (I.9). I concluded that many if not all forms of intentionality are systematically dependent on the kind of intersubjective practice of ascription delineated in I.7 and excluded those which potentially do not from my present analysis.

II. Intentionality in Cognitive Neuroscience

II.1. Representations in the Cognitive and Neurosciences

In psychology and the cognitive sciences, "cognition" usually means information processing related to psychological functions (see e.g. Anderson 2009: introduction). For example, if one of our cognitive abilities is to react aversively to snakes, then the respective function is fulfilled by a process associating an informative representational input, namely perceptions of snake-like things, with the appropriate output, namely aversive behaviour. We can tell the same story about higher cognitive functions such as the ability to identify correctly formed English sentences: here, the respective function is also fulfilled by associating an input, namely perceptions of sentence-like structures, with the appropriate output, namely corresponding judgments of correctness or incorrectness (see Levine 1987: 250). Properties such as representing, carrying information or generally "being about" something are called *semantic*, and what representations represent or information informs us about is correspondingly called *semantic content* (in this context, sometimes also "mental content" or just "content", compare section I.4.1). In our example, the primary pieces of information involved are that what is perceived is a candidate for being an English sentence, like a string of words or characteristic phonemes, and that it is either correct or incorrect.

The notion of representation is central to the cognitive sciences. For instance, according to Thagard,

> "the central hypothesis of cognitive science [is this]: Thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures. Although there is much disagreement about the nature of the representations and computations that constitute thinking, the central hypothesis is general enough to encompass the current range of thinking in cognitive science, including connectionist theories" (Thagard 2005: 10). "Without a doubt, (…) [this hypothesis] has been the most theoretically and experimentally successful approach to mind ever developed. Not everyone in the cognitive science disciplines agrees with [it] (...), but inspection of the leading journals in psychology and other fields reveals that (…) [it] is currently the dominant approach to cognitive science" (ibid.: 11).

One branch of cognitive research is cognitive neuroscience: the investigation of how the brain's properties relate to or underlie cognition (cf. Sullivan 2015b: tba). In humans and many animals, the brain is a prerequisite (or the "basis") for cognition, insofar as its activity is necessary for cognition to occur. This view is commonplace in the cognitive sciences and

recent philosophy of mind: "mental phenomena are biologically based: they are both caused by the operations of the brain and realised in the structure of the brain" (Searle 1983: ix; see also Davidson 2004: 180 and Cummins 2000: 133).

Three major sources of evidence for this assumption are, firstly, lesion studies, where the severe damage to or the complete lack of a certain brain region (or in general, a neural pathology that impairs the activity of certain brain areas) is associated with the failure to perform certain cognitive functions: some "evidence about brain functioning is gathered by observing the performance of people whose brains have been damaged in identifiable ways. A stroke, for example, in a part of the brain dedicated to language can produce deficits such as the inability to utter sentences" (Thagard 2005: 9). Secondly, neuroimaging studies suggest that certain brain activity is systematically correlated with cognitive performance, or as Haynes and Rees put it very generally, "many human neuroimaging studies have provided strong evidence for a close link between the mind and the brain" (Haynes & Rees 2006: 523; for more information both on lesion studies and functional correlations see D'Esposito & Wills 2000). Furthermore, fulfilling psychological functions consists of properly relating environmental cues, internal states and behaviour: If I am asked to recognise a larch from quite a distance, then my perception of the faraway larch, my memory of what larches look like, my estimation of what larches would look like from quite a distance, and my behaviour, signalling my recognition of the larch, need to be properly related. In all instances, these relations require or consist of cognitive processing; so the fact that, thirdly, the brain's physiognomy and activity is what crucially connects and regulates perception, internal bodily states and behaviour also singles it out as the prime candidate for the basis of cognition. For these three reasons, I will take the brain's proper functioning to be a necessity for paradigmatically cognitive functions.

In the first chapter we have seen how some forms of psychological analysis are intertwined with the notions of information, representation and content. Holyoak goes so far as to view these notions as some of the field's defining features: "Psychology is the science that investigates the representation and processing of information by complex organisms" (in Wilson & Keil 1999: xxxix). But how does semantic content figure in cognitive neuroscience? Curiously, just like Holyoak does in the case of psychology, Albright and Neville emphasise that "cognitive neuroscience is (…) a science of information processing" (in Wilson & Keil 1999: li). Now, since insights about the brain's role in cognition are often obtained by correlating brain activity with performance during psychological tests, we know that at least due to this significant methodological overlap cognitive neuroscience inherits one

notion of representation from psychology (again, see D'Esposito & Wills 2000, but also Gold & Stoljar 1999). However, the claim that cognitive neuroscience is a science of information processing would be rather trivial, and much less of a defining characteristic, if it were exclusively true due to the fact that it is intertwined with *another* field about information processing, namely psychology. Rather, what I take Albright & Neville's claim to mean is that cognitive neuroscience is the science of *information processing in the brain*, in the sense that neural properties are justifiably and truthfully describable as having representational features. So, our concern in this chapter is not with how descriptions in cognitive neuroscience can inherit the psychological notion of representation, but rather, how the brain itself can properly be described as carrying semantic features. Cognitive neuroscience's relation to psychology is what may lend us the methodology and the theoretical framework to identify neural properties *as* carrying information, but ultimately, what interests us is how *neurobiological features* themselves get to have semantic properties. Thus, the principal question I will be concerned with is not *how* representational features in the brain are to be identified – although I will say a few things about this issue as well – but *why* some features of the brain are representational in the first place.

II.2. Are Representations at Odds with Naturalism?

In order to see how the notion of representation connects with neuroscience, it is helpful to distinguish between neurobiology and cognitive neuroscience the way Gold and Stoljar do:

> "According to one conception of neuroscience, perhaps the more traditional conception, neuroscience is to be understood as the science we will call *biological neuroscience,* the concern of which is the investigation of the structure and function of individual neurons, neuronal ensembles, and neuronal structures. For simplicity, we will stipulate that biological neuroscience includes only neurophysiology, neuroanatomy, and neurochemistry, and we will take it to be synonymous with *neurobiology.*
>
> According to another conception, neuroscience is taken to be what is often called *cognitive neuroscience* (see Gazzaniga 1995; see also Kosslyn & Andersen 1992 and Kosslyn & Koenig 1995) (...). Cognitive neuroscience is an interdisciplinary approach to the study of the mind, the concern of which is the integration of the biological and physical sciences – including in particular biological neuroscience – with the psychological sciences to provide an explanation

> of mental phenomena. Although biological neuroscience is interested in understanding the biology of the brain, cognitive neuroscience attempts to synthesize biology and psychology to understand the mind. Cognitive neuroscience therefore includes biological neuroscience as a proper part but is not exhausted by it" (Gold & Stoljar 1999: 813).[62]

In spite of its carrying biology in its name rather than physics, neurobiology is ultimately couched in physicalist theories: the brain itself is expected to be wholly describable in physical terms (such as electrical properties or physical properties of neurons) and chemical terms (such as the chemical features of neuro-transmitters) (cf. Griffin & Baron-Cohen 2002: 104). Combined with my earlier remarks, the question now becomes: how do some physical objects, namely brains (or some of their spatiotemporal parts), get to have representational content? Not only does our taking them as representational follow from interpreting Albright and Neville's definition in a non-trivial sense, but *calling* them this way is standard in the scientific literature (see the following section for some examples). So the initial conundrum is that we take physical objects to be representational, while being representational is itself not a physical feature: "A natural definition of representational content only refers to concepts of the natural sciences, which themselves neither are intentional nor draw on the interest of an external observer"[63] and "a naturalistic theory of semantics [is one] where representations, their content, their truth or falsehood are defined without recurrence to terms which themselves are already intentional" (Zehetleitner & Schönbrodt 2013: 197).

Seeking to integrate the notion of representation into the natural sciences in general or neurobiology specifically, we are faced with the challenge of naturalising representational descriptions, i.e. coming up with alternative (coreferential/coextensive/coexplanatory) descriptions which contain nothing but physical terms. These attempts have proven problematic for various reasons (see e.g. Fodor 1974). However, variants of naturalism which are not strictly physicalistic have proven more fruitful, insofar as they have pointed out strategies for swapping representational terms for non-representational terms, such as those from biology (cf. Dretske 1981, Millikan 1989, Zehetleitner & Schönbrodt 2013) or dynamic systems theory (cf. Bischof & Zehetleitner 2015, Zehetleitner forthcoming; see section II.7.2). Such naturalistic programs are not strictly physicalistic, insofar as they use teleological terms (see I.4.5 and Braillard & Malaterre 2015: 9-15) such as "function" or "organism" which are

[62] I take Gold & Stoljar to mean that neurobiology's methodology and conceptual inventory is a proper part of cognitive neuroscience, but not its entire domain. That is, neurobiology can also say something about brain parts which are not cognitively relevant (such as those partaking in internal bodily regulation).

[63] Quoted from a presentation held by Michael Zehetleitner at the LMU's Research Center for Neurophilosophy and Ethics of Neuroscience on Oct 22nd 2013. A similar phrasing can be found in Zehetleitner & Schönbrodt 2015 on p. 197.

not a proper part of physics (which does, however, not exclude the possibility that such theories will prove reducible to future physicalistic theories.[64] However, see Sullivan 2009: 518 for some of the problems reductionism currently faces in neuroscience).

In the following, I will take non-physicalist naturalism to provide sufficient criteria for what counts as scientific analyses of representational properties. Compared to physicalism, this non-physicalist naturalism requires little more than accepting that biological kinds are respectable scientific terms, and that accepting teleological, functional notions into scientific theories does not amount to heresy. I will certainly not attempt to show how representations reduce to strictly physical terms, but rather that they can still be cashed out in thoroughly scientific terminology and that proper cognitive sciences need not be afraid of them. My point is not that there *cannot* be physical descriptions of representations, but rather that, *even if* it turns out that they cannot or will not be reduced, physical theories by themselves do not provide the kind of explanation we seek in the cognitive sciences. That is, perhaps we *will* arrive at physical descriptions of, say, psychological kinds; but since physics does not have any nomological use for such objects, as psychology would, they would cease to be such kinds and thus cease to explain what they were meant to explain (cf. Fodor 1974 and section II.8.4.4). It is far more likely that these redescriptions of non-physical terms in physical terms would create "big data" that is hard to cash out as an explanatory surplus; and whenever big data about psychological properties does explain something, it usually does so by algorithms which themselves do not rely exclusively on physics (– How does facebook predict our behaviour? By statistically correlating certain biographical properties with other biographical properties; see section I.6.5).

II.3. Neural Representations are Sparse

What is it that makes some physical object a carrier of information? And what exactly are these "semantic properties" which we assume neural features to have? First off, of course

[64] Coming out of early 20th century positivism, scientific naturalism has often been taken to amount to physicalistic naturalism, and reducibility to physics as a criterion for what counts as a respectable scientific theory (see e.g. Carnap 1931; for a weaker version, namely reducibility as an empirical "working hypothesis", see Oppenheim & Putnam 1958). However, this physicalist optimism has been waning over the past 100 years (see, again, Fodor 1974). While it is generally accepted that laws and natural kinds in biology and psychology are less strict than in physics (see section I.6.6), this realisation has led to widening the scope of what counts as scientific rather than depriving everything less strict than physics the status of being a science. (Also, our picture of physics itself has changed considerably over the past century, contributing to this widening.) Of course, there is much more to say about the successes and failures of physicalism and its relation to cognitive science, but that is beyond the scope of this book.

no one expects neural representations to be straightforwardly like other forms of symbolic representations familiar to us (see I.4.2): we are not going to literally find signs, pictures or words in the brain, and we should not expect anything in the brain to have representational features outrightly similar to these.[65] For example, if something green is perceived, then the corresponding representation within the brain will of course not be a green picture – and not just because "[b]rain processes are not the sort of things to which colour concepts can be properly applied" (Place 2002, 59). Firstly, for X to represent Y, X is generally not required to share any characteristic property with Y (compare Danto 1981: chapter 1).[66] Representing a meadow does not require a green picture (just take a look at any of Van Gogh's many non-green representations of nature). For this reason alone, imagining anything green does not require our cognitive apparatus to have any green properties.

This is not to say that similarity cannot play a role in an object's representational features: We often do decipher an image intuitively if it looks like what it depicts. The point is rather that similarity can only ever explain a small part of what makes a representation representational (also see I.4.2). It explains a part of pictorial representation (namely those images which look like what they depict), but another significant part of it, such as abstract art or signs, remains unexplained. Linguistic representation, which does not rely on similarity at all, also remains grossly unexplained. And the form of representation we're presently concerned with would remain unexplained as well, since the only instance in which a neural representation is substantially similar to what it represents is when it carries information about neurons. (Sometimes, neuroscientists may even think about the neurons they're currently using to think about these neurons.) In the cases of pictorial and especially linguistic representation, the explanatory gap can be filled with theories about association, learning or convention: if we do not immediately see what a word or sign is meant to represent, we can learn to associate one with the other – which is one of the most important cognitive abilities underlying pictorial or linguistic representations. However, no such process can give physical properties of brains representational qualities: just as absurd as the expectation to find anything green in the brain is the idea that something in the brain represents anything by

[65] Perhaps under one of its more implausible interpretations, Fodor's "language of thought"-hypothesis could indeed be said to assume word-like entities floating around in the brain. But even so, these entities would not be words in any straightforward sense. Rather, as I have argued in I.4.4, LOT should more plausibly be interpreted as being true if it turns out that neural processes underlying cognition satisfy certain (syntactical) requirements which languages generally satisfy. For instance, they should be interpretable as implementing a formal structure which supports compositionality. As I have also pointed out, this interpretation runs the danger of making LOT compatible with connectionism, but that is the price I believe we must pay for plausibility.

[66] Compare also Danto's second chapter, in which he introduces a striking analogy to the theory of action, which underlies intentional psychology. The basic analogy, briefly: Two identical physical objects can radically differ in meaning (i.e. in their roles as signifiers), and so can two identical instances of behaviour (also compare Danto 1973: ix f.).

convention (because there are no homunculi, no little agent-like objects in the brain, for whom these conventions would be significant; cf. Kenny 1971: 65 f. and Levine 1987: 254).

Thus, far more illuminating than the first reason why we should not expect green pictures in the brain, namely that they needn't be green in order to represent green, is the second one: that neural representations are *not symbolic*, insofar as they rest neither on similarity nor convention. This point may seem obvious enough, but many theoretical problems dealing with representations in the cognitive sciences (including many issues about the relations between mental states and brain states) are tracable back to this simple mix-up. When we find someone bemoaning the fact that rationality and normativity, aspects which are inherent to matters of representations as understood in intentional psychology, are a stumbling block for the cognitive sciences, we must assume they have mixed up one form of intentionality for the other. I have distinguished between rich and sparse notions of intentionality in section I.4.4, and none of said problems come with sparse notions (compare footnote 101). Which does not mean that an account of how sparse notions can support semantic ascriptions of information is trivial (an account which I am about to develop in the following sections) – it just means that there are several distinct problems associated with intentionality, and just saying that we are dealing with representations does not imply that we have to deal with all of them at once.

For example, let's assume that there's a pattern of activation in the brain's fusiform face area (FFA) which represents a certain face (cf. Kanwisher 2001). Clearly, one does not represent the other by way of convention: it has never simply been conventionally decided to associate one with the other. Sure, conventions can have a significant impact on neural processing: The fact that we associate the word "bridge" with any particular bridge is by way of convention; and associating the word "bridge" with the perception, memory, sound and function (etc.) of an actual bridge is thanks to neural processing. Thus, representing something symbolically requires shaping our neural processing. (Which is not saying much more than that everytime we learn how to use a symbol, our brain has to follow suit.) Of course, none of this implies that whatever in the brain represents bridges does so by convention. The difference being that our learning about the fact that there is a conventional association between the word "bridge" and actual bridges causes the brain to rewire itself so as to fulfill the associated cognitive function. Thus, said linguistic convention that "bridge" means bridges is a *cause* for something happening on the neural level, but whatever happens on the neural level does not itself represent by convention.

To drive this point home, compare Cummins' theory of psychological explanation: Analogously to what I have said thus far, he takes psychological capacities to be characterised by carrying out information-processing (see Levine 1987: 250 f.). According to his account, whatever device implements such an information-processing programme (read: the brain) need not have access to a *representation* of the programme's instructions it is carrying out. What matters is rather that said device's causal structure is so organised as to yield the desired (correct or adequate) output dependent on a given input. The representation itself is not a separate element of the causal stream between in- and output (ibid.: 256). And yet, Fodor seems to disagree: "What distinguishes what organisms do from what [non-cognisers] do is that a *representation of the rules they follow constitutes one of the causal determinants of their behaviour*" (Fodor 1975: 74, fn. 15, author's emphasis). However, both Cummins' and Fodor's points are right on target and readily reconcilable: Namely, Fodor's point is that a representation of the rules, say, a textbook of the English language, is one of the causal determinants of an English-speaking student's linguistic behaviour (compare section I.4.4). So, if this student has picked up English by a textbook, then the textbook is both a representation of the rules the student follows as well as a causal determinant of her behaviour (cf. Levine 1987: 259). The same goes for everyone who has ever learned anything: Who- or whatever has been key in teaching them is a causal determinant of their learned behaviour.

Now, Cummins' point is that the device which allows the learner to carry out the learned program – in this case our brain, which allows us to correctly form and react to English sentences – does itself not consult an "internal textbook" or the like while producing English sentences. Insinuating that there's a tiny "homunculus [in the brain] pulling a volume off the shelf" (ibid.: 254) whenever we produce these sentences is willfully misleading. Rather, the brain's part in allowing us to speak English consists in causal neural sequences – causal sequences which ultimately owe their manifestation to external textbooks (or similar sources of learning).[67] This is what it means to say that external representations and conventions can have causal impacts on neural structures, while at the same time, there are neither textbook-like nor conventional representations in the brain.[68]

[67] This is what Paul Churchland would call "third-level learning", i.e. cultural learning depending on communication (cf. Churchland 2012: chapter 5), rather than first-level learning, which is described as the shaping of neural networks through gradual alteration of synaptic weights between neurons and Hebbian plasticity (ibid.: chapter 2).

[68] Here, Cummins uses the term "representation" more restrictive than I do by only applying it to things like textbooks and not to neural structures. But this is only a terminological issue: Of course I agree that if textbooks are the paradigmatic example for representations, then there is no such thing in neural form. However, I am using the term "neural representations" in a distinctly defined technical sense which is meant to encompass things radically different from textbooks, and I have already stated that neural representations do not represent by way of convention. (However, it should be noted that Cummins does not generally use the term in this sense, compare footnote 96).

So, the fallacious idea of finding pictures (or the like) in the brain is rooted in mistaking sparse for rich notions of representations. What we find in neurobiology are sparse notions of representations to whom intentionality is of no integral explanatory use. In other words, semantically individuated kinds are not neurobiological kinds (at least not in virtue of their semantic individuation). On the most basic level, specific non-representational causal notions do the explanatory work (see footnote 51). Such a sparse notion of representation goes like this:

> "[w]hen the firing pattern of a neuron is significantly correlated with the presence of some feature of a stimulus that an organism is currently experiencing, that pattern is said to represent that feature. [It is assumed] that the relationship between a neural representation and whatever it is about is causal: a neuron will only exhibit a significant increase in its firing rate above baseline in response to that stimulus feature that causes it to fire. Whatever stimulus feature causes it to fire in this way, it represents (...). Support for this assumption comes from cognitive neurophysiological investigations of predominantly sensory neurons. For example, neurons in auditory cortex fire in response to auditory stimuli, neurons in insular cortex to taste stimuli, and so forth for other sensory systems" (Sullivan 2010: 876 f.; for the technical details see Dayan & Abbott 2001: chapter 10).[69]

If we take this as a definition, "neural pattern A represents X" should be taken as synonymous with "A is significantly correlated with the presence of X" and/or "A is caused by X". While we have seen that sometimes representational properties are established by way of causal linkage (see section I.4.3), a causal relationship between two objects is by itself not sufficient for establishing that the effect semantically represents the cause. (Sullivan is still correct, because it is true that neuroscientists treat neural events or processes which are significantly correlated to semantic properties or stand in a causal relation to these as representations and that they are justified in doing so. However, it is wrong to say that they are justified in doing so merely *because* these are correlated in this way or stand in said causal relationship.) If the causal relationship between neurons and what they are said to represent were all that connected them, this notion of representing would be sparse indeed: It would be so sparse that semantic notions could be completely discarded in favour of non-representational causal explanations. If the intentional explanation yields no surplus at all, then ascribing intentional properties to neural structures would not be justifiable. And if that were the whole story there is to tell about representation, then we could just scratch the

[69] This notion that an organism O represents X as R if X causes R in O can be traced back to *Lockean Covariance* (see Cummins 1991: ch. 4). Also compare Block's writings on *Correlationism* in his 2007: 485-487.

problem of intentionality in neurobiology off our list. However, as I am going to show in the next section, this is not the case.

For now, consider some other examples, which are all paradigmatic for the use of the term "representation" in neuroscience:

- Canonical neurons represent affordances of objects, such as a cup's being suited for being handled with a precision grip (cf. Grèzes et al. 2003).
- Mirror neurons are neurons which fire both when an action is performed and when the same action is observed. This mechanism has been hypothesised to represent other people's actions, their intentions when performing them, their emotions (related to the means of their expression), and/or other mental states (cf. Di Pellegrino et al. 1992, Iacoboni et al. 1999).
- The cortical homunculi represent body areas: one related to somatosensory properties, another to motor properties (cf. Penfield & Rasmussen 1950).
- Internal maps: "Place cells are said to 'represent' (...) locations and are taken to play a role in the formation of 'cognitive or spatial maps' (...). Spatial maps are thought to be representations that are distributed across place cells, with each place cell contributing that aspect of the environment it represents to the map" (Sullivan 2010: 879).

All of these forms of neural representation, on the face of it, do not require representational notions to analyse their functioning. Therefore, such analyses are sparse. To reiterate: A representation is sparse if it can be analysed without recurring to any form of symbolic representation.[70] Such analyses can be readily supplied for our examples: Canonical neurons are an integral part of the following mechanism (and others like it): when perceiving a tool whose handling requires a precision grip, the motor functions that enable us to execute a precision grip are activated automatically by perceiving this object. In the case of mirror neurons, "representing another person's emotion internally" is shorthand for "perceiving another person's emotional state activates a functional equivalent to the other person's

[70] If you happen to believe that symbolic representation can itself be reduced to non-symbolic representation, then the distinction goes like this: Representation is sparse if its analysis does not require invoking what symbolic representation characteristically reduces to. Even reductionists will have to concede that symbolic representation requires more than sparse representations such as those invoked in these examples, even if this "more" does not consist in irreducible semantic properties. That is, if semantic properties reduce to causal relations, they still reduce to a specific set of causal relations, not to all of them; and thus, being a causal relation is not sufficient even to pick out reducible semantic properties. Rather, this subset of causal properties has to be marked as being what semantic properties reduce to, and this need for marking means that semantic properties are real and explanatory.

neuronal state underlying this emotion, which is isomorph with the neuronal state my own brain would display if I were in that emotional state". Saying that the cortical homunculi represent (parts of) the body means that there is a certain amount of neurons in the primary motor cortex and the primary somatosensory cortex dedicated to (non-representationally) processing the relevant properties of each corresponding body-part.[71]

It's harder to come up with a concise non-representational rephrasing of the way internal maps work, because when asked to specify what those neurons whose activity underlies cognitive maps actually do, we tend to summarise what actual maps enable us to do. In fact, the reason why they are called "internal maps" in the first place is because internal maps enable us to do what we usually associate with what maps enable us to do. However, the fact that we can orient ourselves using external maps depends on a rich notion of intentionality, insofar as it depends on our understanding that the map (conventionally) represents the respective area. That is, the non-intentional properties of the map alone are not sufficient for it to function as a map: it needs to be embedded into a symbolic practice which uses physical objects such as maps to signify geographical properties. Nothing of this sort is true for internal maps: the physical properties of internal maps alone are what makes them implement the orientation function. So, while specific properties of some hippocampal cells are in important ways analogous to properties which are represented by maps, it would be fallacious to hold that this analogy establishes their being representational. If, say, the relative strength of the connections between the cells is the same ratio as the relative distance between the spots in which the respective cells are active, then the important analogy consists in the ratio, not in any intentional property.[72] In a nutshell, internal maps do serve similar functions as looking up an external map does, but the brain certainly does not "look up" an internal map in the way we look up external maps.

Note that we do not merely concentrate on the *causes* of neural representations in order to determine whether and what they represent, but also on their *effects*: we also invoke knowledge (or assumptions and hypotheses) about what type of behaviour or cognitive output

[71] Note that the homunculi stand out from the other examples by being pictorial representations of anatomical divisions in the brain to begin with; thus, they are to some degree entangled with symbolic representation. That is, the motor homunculus is a representation in much the same way as a schematic depicting the mechanical relations between a toy car's remote control and the toy car itself is.

[72] On a side note, the functionality and representational quality of external maps is entangled in a way which differs from other forms of representation: For example, if the ratio of the distances between points on the map and the distances between the represented geographical points is not the same, then the map cannot be said to accurately represent these points. No such thing applies to pictorial or linguistic representation. In this way, maps are more accurately characterised as akin to speedometers, which I briefly discussed in section I.4.2. For this reason, functional analyses should be expected to reveal important analogies between internal and external maps, even though orientation using internal maps does not require intentional capacities, whereas orientation using external maps does.

is associated with the firing of these neurons. For example, neural representations of faces in the FFA are what causes our recognitional behaviour. Harris et al. do much the same in their stipulation of what beliefs are on a neuronal level: "The capacity of the human brain to believe or disbelieve ostensible statements of fact (eg, 'You left your wallet on the bar.' 'That white powder is anthrax.') is clearly part of its machinery for the initiation and control of complex behaviour" (Harris et al. 2008: 141). Here, they take the operationalisation of the notion of representation even further by stating what causal role specific forms of intentional states, such as beliefs, play when framed in a neural context.

Considering that ascriptive practice plays such an important part in intentional psychology, let me add some cautionary remarks. An operationalisation along said lines needs to be distinguished from common folk-psychological practice, otherwise it would be subject to what Bennett & Hacker call the mereological fallacy. According to them, brains don't think just as stomachs don't eat – the respective predicates apply to whole persons only, but not to their parts (cf. Bennett & Hacker 2003: chapter 3). In Harris-type examples, however, its use is justified for two reasons: On the one hand, the neuronal firing correlates to the intentional states, suggesting a systematic connection, and on the other, neural activity is instrumental in causing behaviour that provides evidence for assigning intentional states. That is, if the FFA provides causal grounds for face-recognitional behaviour, then we are justified in saying that the FFA plays an important part in recognising faces. Prima facie, this violates Bennett's and Hacker's criteria, on whose view the assignment of intentional states is exclusively justified on the grounds of common ascriptive practice (ibid.: chapter 3.9; see also section I.7), and it would certainly be wrong to insinuate that the grounds on which we judge the FFA to represent faces are analogous to those on which we commonly assign intentional mental states. But of course calling some neural activity representational is meant to insinuate no such thing. It may occasionally cause confusion for the layman, but the surplus for theories in the cognitive sciences more than justifies adapting the notion of representation for patterns in the auditory cortex, the FFA and the like. As Dennett has pointed out, while there is a distinction between concepts we use at the personal and subpersonal levels (cf. Wittgenstein 1953: §281), this point

> "has occasionally been misconstrued (…) as the lesson that the personal level of explanation is the only level of explanation when the subject matter is human minds and actions. (…) [Rather, t]he recognition that there are two levels of explanation gives birth to the burden of relating them, and this is a task that is not outside the philosopher's province. (…) There

> remains the question of how each bit of the *talk* about pain is related to neural impulses or talk about neural impulses" (Dennett 2007: 79).

Furthermore, to say that theories about neural activity stand in an explanatory relation to mental states, or behaviour which amounts to evidence for the ascription of mental states, is not to commit to the claim that these theories can fulfill the same roles as kinds from intentional psychology. For example, while we can imagine there being different explanations of the same behaviour, one psychological, one neuronal, the explanations serve different roles. Consider that one explanation might read "Sam got angry because he saw Max being mistreated unjustly" [E_1], while another might read "Sam got angry because his amygdala was stimulated" [E_2]. Let's assume that, in their respective fields, these are valid explanations for the same state of angriness which Sam experiences. Now, E_1 and E_2 can be differently illuminating. What follows from E_1 is that, if Sam's anger is righteous, then we should see to it that Max's unjust treatment is rectified.[73] However, if Sam's anger is misdirected, E_1 implies that we should see to it that Sam realises that Max's treatment was in fact justified (or that Max was not mistreated at all). On the other hand, if Sam's belief is not rooted in an actual state-of-affairs, that is, if he hallucinated, or suffers from an illness that causes him to see people being treated unjustly, then we might consider a therapeutical intervention. All of this follows from E_1, while nothing of the sort follows from E_2. Yet, E_2 provides us knowledge necessary to medically intervene in the latter case or insight into how to build AIs which could simulate angriness by way of neural networks, or the like.

II.4. Encoded Information, Mindreading and Correlations

While in the previous section I have shown that causes and effects of neural structures are characteristic for their being treated as representations, I am also going to argue that there is far more to neural representation than that. A first step on this path consists in pointing out that the notion of neural representation is not merely grounded in the notion of causality, but also in the assumption that such representations carry *encoded information*. So, what we should really expect to find in the brain is neither symbolic representation nor mere correlates of external semantic properties, but genuine information which is encoded in the brain's properties and its activity.

[73] Analogously, matters of irrationality are only applicable given an intentional explanation (cf. Davidson 2004: 180).

Usually, the idea of information being encoded means that someone came up with an encoder which translates one set of representations to another, according to a set of fixed rules. Of course, no such thing is true for the brain, since neither did an agent create its neural code, nor does the brain consult rules (see the previous section). Rather, saying that information is encoded in the brain assumes a technical use of the term: it means that there is a translational algorithm by which this information could be extracted. That there "is" such an algorithm implies that the neural encoding has a certain property which makes it in principle decodable, not that there actually (or currently) exists some method of decoding it. So the algorithm exists in an abstract sense, not necessarily in an actual sense. Consequently, said property is abstract in nature: it is the property of *correlating with semantic content*.

As far as the actual existence of such algorithms goes, we should look to the ongoing development of "brain decoders", in virtue of which some specifics of neurally encoded information have been uncovered. The decoding method has come out of the development of brain-computer interfaces (BCI):

> "A typical BCI setup is as follows: EEG electrodes are fixed to the patient's scalp. Potential differences due to electrical currents in the brain, originating from neural activity, are fed through an amplifier and into a computer. Algorithms are trained to recognise two conditions, such as imagined hand or foot movement, by repeated recordings (trials) of such imagination tasks done by the patient. This classification of two conditions allows the patient to choose letters or other elements on a computer screen, thus enabling communication between the patient and the outside world."[74] "Such voluntarily controlled brain signals can subsequently be used to control artificial devices to allow subjects to spell words or move cursors on computer displays in two dimensions. Interestingly, subjects can even learn to regulate signals recorded using functional MRI in real-time. It might be possible to achieve even better decoding when electrodes are directly implanted into the brain, which is possible in monkeys (…) and occasionally also in human patients. Not only motor commands but also perception can, in principle, be decoded from the spiking activity of single neurons in humans and animals. However, such invasive techniques necessarily involve surgical implantation of electrodes that is not feasible at present for use in healthy human participants" (Haynes & Rees 2006.: 524).

Using non-invasive methods of functional imaging, it has recently been found that some neural activity allows semantic content to be extracted from it by being measurably

[74] Quoted from the website of Tübingen University: http://www.ti.uni-tuebingen.de/BCI.856.0.html?&L=1, retrieved on May 14th 2013.

correlated with it (see Thirion et al. 2006, Kay et al. 2008, Naselaris et al. 2009, Nishimoto et al. 2011). Coming up with these algorithmic decoders is instrumental in investigating which information is stored where and when, and whether some specific information is stored in the brain at all – as opposed to being "enacted", i.e. being a result of the interaction between an organism's cognitive apparatus and its environment (compare Haugeland 1995). Such "mindreading" decoders also come with the hopes of uncovering covert attitudes, just like lie detectors would (see Haynes & Rees 2006: 528 f.), and of improving said brain-machine interfaces. One major development in recent mindreading algorithms consisted in successfully decoding the information acquired through the perception of natural images (as documented in Nishimoto et al. 2011):

> "Natural scenes pose an even harder challenge to the decoding of perception. They are both dynamic and have added complexities compared to the simplified and highly controlled stimuli used in most experiments. For example, natural visual scenes typically contain not just one but many objects that can appear, move and disappear independently. Under natural viewing conditions, individuals typically do not fixate a central fixation spot but freely move their eyes to scan specific paths. This creates a particular problem for decoding spatially organised patterns from activity in retinotopic maps, as eye movements will create dynamic spatial shifts in such activity" (Haynes & Rees 2006: 527).

The fact that certain neuronal firing patterns are correlated to processed semantic content, which is the methodological means for extracting information from the brain, can be traced back to a systematic relation between semantically characterised functions and their neural implementation: Say, let's assume that some salient features of a given environment provide an agent with information which we know she can extract from these (such as her watching movies about aeroplanes; see Nishimoto et al. 2011: 1644). For reasons stated in section II.1, we should also assume that it is her brain which enables her to do so by initiating neural activity – activity which, if it is systematic enough, should be correlatable with the information the agent extracts from this environment. So, by identifying what kind of neural activity needs to feed into a decoding algorithm, we are provided with information about the neural implementation of a behaviourally observable function. Since the execution of the respective cognitive function was verifiable behaviourally to begin with (i.e. there is behavioural evidence for judging whether the person in question gains the knowledge that a given movie is about aeroplanes from watching it), and since we can assume that there needs to be a systematic neural cause for this kind of behaviour, we can expect that, using the right

methods, we can find neural activity which is correlated with the retrieved information. That is, if an agent can retrieve said information and if we can assume that this retrieval is directly caused by neural activity (and indirectly by the original source of information, in this case a movie), then behavioural expressions of retrieval (such as saying "the movie featured aeroplanes") must be realised by motor activity initiated by specific neural firing. Interpretating our correlations causally, we can hypothesise that the "encoded information", i.e. the neural activity feeding into the algorithm, is efficacious in the process leading up to the behaviour which counts as retrieval. We can err by mistakenly assuming coincidental correlations to reveal causal efficacy, but we methodologically assume that there needs to be a causal neural process which leads from an agent's watching a movie about aeroplanes to behavioural expressions such as saying "the movie features aeroplanes". In this way, in- and outputs *must* be neurally connected, and tracing this connection as an at least temporally and individually localised correlation between semantic properties and neural activity then depends on the quality of the employed method.

No formal requirements are placed on this encoded information other than that it stands in a causal relationship to the pre-encoded information (such as the content of the movie) and that it specifically leads up to its retrieval: This process need not be modelled as a deductive implication or as a deterministic causal sequence in order to arrive at the conclusion that it will still correlate with the information that is retrieved. Even if it turns out that the relevant parts of the brain were to operate rather chaotically and would only ever yield statistical results, it would still suffice. The only requirement, as Haynes & Rees concisely state, is this: "In theory, if the responses at any brain location differ between two mental states, then it should be possible to use measurements of activity at that brain location to determine which one of those two mental states currently reflects the thinking of the individual. In practice it is often difficult (although not always impossible) to find individual locations where the differences between conditions are sufficiently large to allow for efficient decoding" (Haynes & Rees: 523). So, we need not generally decide at what level brain properties are supposed to be "read" – at the level of fine-grained neural activity, fMRI-data, or anything inbetween –, since what matters is merely that what is used as an input for the decoder supports an "inverse inference" (Thirion et al. 2006: 1104) from neural to mental state. Note that this requirement allows for mindreading to have been accomplished when what is in fact "read" is only a good enough indicator for the respective mental state, and not the mental state itself. For example, it may be the case that the neural correlate of a folk song is what is activated and effectively detected in fMRI when thinking of a singer, so that the

former is a reliable indicator for the latter, even when the respective study boasts about having read the subject's mind *as* thinking of the singer (cf. Beck 2014: 22).

Therefore, the fact that we can find activity within the brain that is correlated to information which we know the agent can extract (such as said movies being about aeroplanes) is more informative of our technological and methodological advancement than regarding a fundamental insight into why our brains have representational capacities: "it should, at least in principle, be possible to decode what an individual is thinking from their brain activity. However, this does not reveal whether such decoding of mental states, or 'brain reading', can be practically achieved with current neuroimaging methods" (Haynes & Rees 2006: 523). So, we should regard recent "mindreading" studies as demonstrating that brain activity can indeed to some degree be decoded using current methods, and that there are hopes of using future methods to improve upon this decoding. As Haynes & Rees note regarding some of the specific methodological problems, which have been successfully circumvented in some recent studies,

> "[m]any detailed object features are represented at a much finer spatial scale in the cortex than the resolution of fMRI. (...) Nevertheless, recent work demonstrates that pattern-based decoding of BOLD contrast fMRI signals acquired at relatively low spatial resolution can successfully predict the perception of such low-level perceptual features (…). For example, the orientation, direction of motion and even perceived colour of a visual stimulus presented to an individual can be predicted by decoding spatially distributed patterns of signals from local regions of the early visual cortex. These spatially distributed response patterns might reflect biased low-resolution sampling by fMRI of slight irregularities in such high resolution feature maps (…). Strikingly, despite the relatively low spatial resolution of conventional fMRI, the decoding of image orientation is possible with high accuracy (…) and even from brief measurements of primary visual cortex (V1) activity" (ibid.: 525).

So, what is essentially up to discovery (and genius of engineering) are the details of the retrieval process; underlying this discovery is the hypothesis that, if we conceptually require outward retrieval behaviour, such as saying "the movie featured aeroplanes", to ultimately be caused by having watched a movie about aeroplanes, and the only things which mediate between watching the movie and retrieving the information are neural properties, then some of these must correlate with the respective information. The crucial insight we gain from mindreading experiments is finding out which firing pattern specifically carries retrievable information (and, once we have a comprehensive view of the brain, this insight

should be continuous with insights into the nature and further steps of the processing which said firing pattern is a constitutive part of). However, what these studies tell us – namely that some of the brain's spatiotemporal parts are correlated with semantic content – does not settle the question we're contemplating, namely what *makes* a spatiotemporal part of the brain representational. As just laid out, the fact that those parts which are representational are correlated with semantic content, and should thusly be decodable, is a basic requirement of their being representational. Still, it is not sufficient: the fact that they are representational is nothing but an assumption at this point, and what we gain from successful mindreading is but the knowledge that there are *correlations* between brain activity and semantic information.

Why is this not sufficient? Well, picture that some brain structures, once they receive input from aeroplane-perceptions, will initiate firing patterns which our mindreading algorithms have been able to significantly correlate with pictures of aeroplanes.[75] As pointed out by Haynes and Rees, the only requirement for this correlation to hold is that the neural activity systematically differs between two mental states. So, what successful mindreading tells us is that the neural effect of a semantic (i.e. information-carrying) cause is (measurably) different from the neural effect of a different semantic cause. This may count as satisfying a physicalist notion of information: namely, saying that if a cause can be reconstructed from its effect, then the effect carries information about the cause (cf. Shannon 1948). Since what is reconstructed in many mindreading studies is essentially information that is present in visual perception, this physicalist notion – i.e. reconstructing the perceptual cause from the neural effect – is the sole requirement for this version of mindreading to work.[76] In other mindreading studies, which are concerned with covert attitudes, causes of or dispositions to actions (such as the infamous Libet et al. 1983), we should be able to tell similar causal stories.

[75] One tacitly assumed requirement for these representations to be triggered is that the cognitive system is in fact ready to process aeroplane-perceptions. That is, certain top-down processes related to attention should not hinder the system from processing aeroplanes, similar to how they could hinder a gorilla to be perceived in Simons' and Chabris' famous test (Simons & Chabris 1999).

[76] At least in principle. Given current imaging methods, which are constrained by low resolution, it may be required to depend on additional areas, which are to a higher degree sensitive to semantic cues than early visual areas: "fMRI data and a structural encoding model are insufficient to support high-quality reconstructions of natural images. (…) However, by applying an additional semantic encoding model that extracts the information present in anterior visual areas, we produce reconstructions that accurately reflect semantic content of the target images as well" (Naselaris et al. 2009: 903). "There is evidence that brain areas in anterior visual cortex encode information that is related to the semantic content of images" (ibid.: 905). "Our results show that the semantic encoding model accurately characterises a set of voxels in anterior visual cortex that are functionally distinct and anatomically separated from the structural voxels located in early visual cortex. The structural voxels in early visual areas encode information about local contrast and texture, while the semantic voxels in anterior portions of lateral occipital and in the AOC encode information related to the semantic content of natural images" (ibid.: 907).

However, knowing that whatever caused a certain firing pattern can be reconstructed from it is far from saying that the firing pattern represents it. Picture a chain of dominos, whose sequential pushing and falling happens in a straightforwardly causal manner, and where falling correlates with being pushed; and where (ceteris paribus) we could even reconstruct information about the force and angle of the cause (push) from knowing its precise effect (fall). Still, any instance of falling certainly does not represent any force or angle in the sense that, say, the fusiform face area represents faces – even though carrying information in the physicalist sense may be (part of) *the method by which* the FFA represents faces. Again, this is because such a correlation is but a *condition* for the representational relation: If a certain firing pattern is the correct neural representation of a salient feature of the environment, such as a movie depicting an aeroplane, then we should expect this pattern to be reliably triggered by the appropriate cue, thus producing correlations. Given that we have reason to assume that neural events won't merely coincidentally be correlated with cognitive processing but track causal processes underlying it, this means that being correlated with semantic properties is an adequate heuristic for identifying neurally implemented representations. But we are only justified in using this heuristic because we already assume that whatever in the brain correlates with aeroplane-perceptions, -memories or -associations is actually representational. So, our real job is finding out the ultimate reason for such an assumption.

As we have seen in section I.4.4, Ned Block also construed the picture of the brain as a syntactic engine driving a semantic engine in terms of a correlation:

> "[I] mentioned a correlation between causal interactions among symbolic structures in our brains and rational relations among the meanings of the symbol structures. This way of speaking can be misleading if it encourages the picture of the neuroscientist opening the brain, just *seeing* the symbols, and then figuring out what they mean. Such a picture inverts the order of discovery, and gives the wrong impression of what makes something a symbol.
>
> The way to discover symbols in the brain is first to map out rational relations among states of mind, and then identify aspects of these states that can be thought of as symbolic in virtue of their functions. Function is what gives a symbol its identity, even the symbols in English orthography, though this can be hard to appreciate because these functions have been rigidified by habit and convention. In reading unfamiliar handwriting, we may notice an unorthodox symbol, someone's weird way of writing a letter of the alphabet. How do we know which letter of the alphabet it is? By its function! Th% function of a symbol is som-%thing on % can appr%ciat% by s%%ing how it app%ars in s%nt%nc%s containing familiar

> words whos% m%anings w% can gu%ss. You will have little trouble figuring out, on this basis, what letter in the last sentence was replaced by '%'" (Block 1995b: 398).

To clarify: What we're dealing with here is not merely the methodological issue that, using current methods, we can *only* discover correlations rather than actual causal relationships (although we should neither deceive ourselves into believing that the latter can ever be identified merely by *looking* – positing causal relationships will usually result from an inference to the best explanation of a given data set, at least in the cases we are dealing with). In the mindreading case, decoders are based on correlations, and we could hypothesise that with finer methods, we might be able to trace the firing patterns which correlate with semantic content to causal sequences rooted in the physicochemical properties of neurons, and thus swap correlations (C) for laws (L):

(C_1) [Perception of movies featuring aeroplanes] correlates with [Neural Firing N_1]
(L_1) [Perception of movies featuring aeroplanes] causes [Neural Firing N_1]

However, this is not the point here. The point is that even if we have good reasons to interpret correlations causally, or even if hypothetically we had some ideal method allowing us to get to actual causal sequences, we would still not gain an answer regarding why these firing patterns actually represent what they do. Compare what's the case with laws of intentional psychology:

(L_{IP1}) [Perception of movies featuring aeroplanes] causes [Belief that the movie features aeroplanes][77]

In this case, the description of the belief alone, namely that the movie features aeroplanes, conceptually implies that the content of the psychological state is "the movie features aeroplanes". However, no such thing is the case for L_1: It does not conceptually follow from any proper description of N_1 alone that it has the content "the movie features aeroplanes", or *any* representational content, for that matter. So, mindreading may establish all sorts of Cs, and future methods may establish all sorts of Ls, but that does not settle our question how neural firing patterns gain their intentionality (see Figure 4).[78]

[77] It goes without saying that C_1, L_1 and L_{IP1} all need appropriate ceteris paribus conditions in order to hold.
[78] Brigitte Falkenburg stresses that "we are dealing with statistical evidence rather than the definitive identification of certain thoughts based on patterns of activity" (my own translation of Falkenburg 2012: 194 f.). Since she is more concerned with matters of jurisdiction (i.e. mindreading as lie detection) than with the

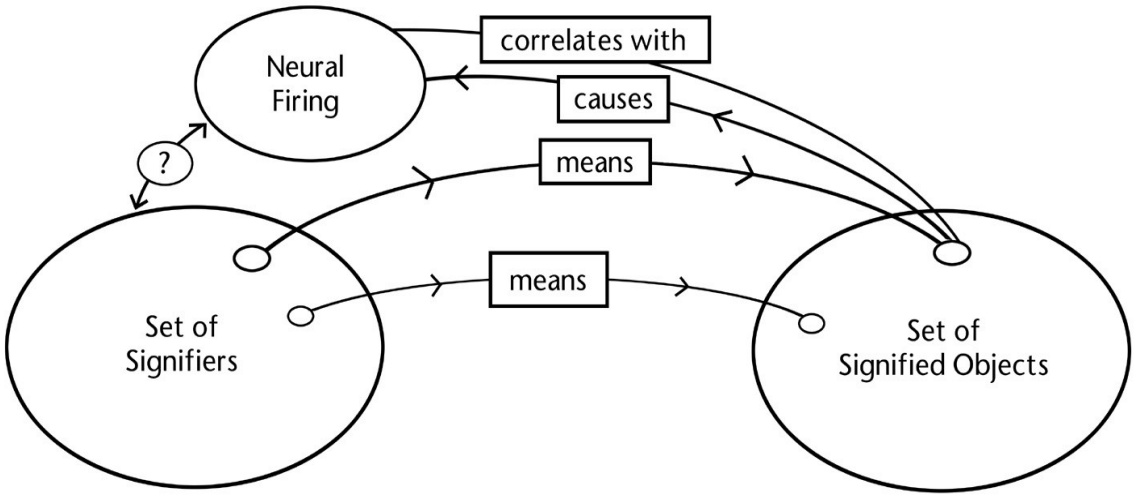

Figure 4. Neural firing correlates with the perceived presence of intentional (i.e. signified) objects and may even be caused by the perception of some, but this does not settle the question whether it's a proper part of the set of meaningful entities (i.e. signifiers).

The point has been made that physical properties can be found to correlate with semantic content, while actually not representing anything – another reason supporting the claim that correlations alone are not sufficient to identify representations. Searle has famously contributed to this debate by stating that

> "[f]or any program there is some sufficiently complex object such that there is some description of the object under which it is implementing the program. Thus for example the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements which is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar then if it is a big enough wall it is implementing any program, including any program implemented in the brain" (Searle 1990: 27).

Many have criticised Searle's point (see e.g. Chalmers 1996, Block 2003, Haugeland 2003) to the effect that mere isomorphism is insufficient for a physical structure to qualify as a computing system. Rather, for something to count as an implementation, it must be a causal process which reliably carries out the specified operations. Yet, what is common to either position is that mere correlation with an information-processing structure is not sufficient for something to count as an information-processing system itself.

---

attribution of mental states, it should be clear that statistical errors will be less tolerable. However, we should keep in mind that psychological attributions only ever apply with a certain probability, based on the quality of evidence at our disposal. So, we should be willing to admit statistical evidence into intentional psychology.

To illustrate this point, picture the patterns male pufferfish create on the seabed in order to attract females (see Figure 5 and Kawase et al. 2013). Certainly, pufferfish create these patterns in virtue of a causal process which reliably carries out specified operations. The similarity of spatial properties between these patterns and the doilies my grandmother used to crochet is sufficient to establish a correlation between them. For example, imagine an eye feeding signals related to visual contrast to a neural network. Whatever the network's activity exactly consist in, as long as it systematically depends on the signals it receives it will correlate in the instances of the eye's scanning the pufferfish pattern and the doily. Still, one activity clearly does not represent the other or carry information which is about the other (even ignoring the additional problem that representation is asymmetric, while correlation is symmetric; crucially, here we find no representational relation in either direction).

While this illustration taps into some intuitions we have about what it means to represent or carry information, there is an additional point to make which goes beyond an appeal to intuitions. (Perhaps neural representation is so novel, idiosyncratic and counterintuitive a notion as to resist such appeals.) Namely, it is the concept of misrepresentation which is integral to matters of representation (cf. Clarke 2004: 50 f., Shope 1999: 279-281, Neander 1995): Since representations have satisfaction conditions – if they truthfully represent their objects, they are true, if not, they are false (cf. Searle 1983: 10) – some causal or covarying effects will have to be marked as "proper" and others as "improper", some as truthfully representing and some as fallaciously representing. But no such concepts are to be found if we merely look at matters of cause or covariance (cf. Cummins 1991: ch. 4-6, Ramsey 2007: 118–150). Sometimes, a toad will mistake a stick for a worm, and sometimes we will see faces in clouds, and any notion of representation that aims to live up to its name has to allow for marking these as instances of error. So, representation cannot be merely cause or covariance tout court; and if it is to be formulated in terms of either, we need to invoke additional means of marking them as proper.

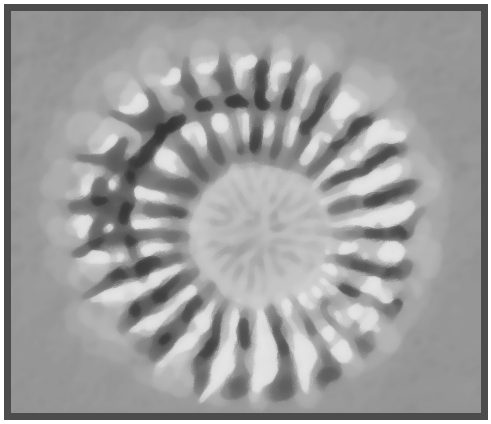

Figure 5: A schematic of patterns created by male pufferfish on the ocean floor, measuring about 2 meters (7 feet) in diameter. See Yoji Ookata's photos at http://ookatayouji.amaminchu.com/archives/2012/09/post_459.html.

So, the argument so far is this:

[P1] We have good reasons to believe that if some spatiotemporal brain-part N_x reliably correlates with content C_x, then N_x represents C_x.
[P2] Correlations between any physical object P and any content C are not sufficient for concluding that P represents C.
[P3] N_x is a P.
[Conclusion] We must have some additional reason for supporting the inference stated in P1 than just its antecedent (namely, that N_1 correlates with C_1).

Our present question can now be sharpened to the point: What additional reason do we have which is needed to infer that if neural firing-patterns are correlated with some content, then they represent this content? If we take certain neural properties to be carriers of information, then this cannot be solely on the grounds that we find them to correlate with the respective content, but because of other facts we know about the brain (or which we at least expect to be true of it). For this reason, I am going to investigate these further facts – namely, some facts going into the general model of scientific explanation in cognitive neuroscience. So, let's look at how research in this area formally proceeds.

II.5. Functional Analyses of Intentional States

II.5.1. The Hard Part of the Easy Problem

David Chalmers famously divided the scientific research of consciousness into a hard and an easy problem. "The easy problems are easy precisely because they concern the explanation of cognitive *abilities* and *functions*" (Chalmers 2010: 6), whereas the hard problem of investigating experience (i.e. qualitative conscious states) goes "*beyond* problems about the performance of functions" (ibid.: 8). Functional explanations might play a role in coming up with an explanation of conscious experience, but any "key insight that allows an explanation of experience (…) will be an *extra* explanatory reward" (ibid.) beyond functional explanations. More than twenty years earlier, Thomas Nagel had already noted: "If we acknowledge that a physical theory of mind must account for the subjective character of

experience, we must admit that no presently available conception gives us a clue how this could be done" (Nagel 1974: 176).[79]

So, the comparatively easy problems – which as Chalmers stresses, are only *comparatively* easy, since "getting the details right will probably take a century or two of difficult empirical work" (Chalmers 2010: 5) – "are those that seem directly susceptible to the standard methods of cognitive science, whereby a phenomenon is explained in terms of computational or neural mechanisms" (ibid.: 4). What is required for their solution is thoroughly analysing the mechanisms which perform cognitive functions. Crucially, these mechanisms are themselves integratable into physicalist theories: Any such mechanism is but an arrangement of physical parts and whose interactions in space and time are governed by the laws of physics.

Mechanistic explanation is one of the key explanatory concepts currently employed by the cognitive sciences, and it essentially connects neuroscience, which in turn integrates sciences such as physics, chemistry and biology, with psychology (see Bechtel & Wright 2009). Here, the basic idea is that internal mechanisms, which are formally or functionally described by any theoretical cognitive science and tested on a behavioural level by experimental psychology, are implemented on a neuronal basis, thus lending "naturalistic" physical grounding to the realm of human cognition.[80] So, if we seek to make headway with the tools currently employed in the cognitive sciences, and if we seek to relate human cognition to fundamentally physicalist theories, mechanistic explanations are a promising way to go.

One particularly interesting "easy problem" Chalmers mentions is "the integration of information by a cognitive system" (Chalmers 2010: 4), and to explain this phenomenon, he says, "we need only exhibit mechanisms by which information is brought together and

[79] I'm not going to go further into what Chalmers calls the "hard problem", but for a striking criticism of separating the empirical investigation of consciousness from its functional construal see Cohen & Dennett 2011: "All theories of consciousness based on the assumption that there are hard and easy problems can never be verified or falsified because it is the products of cognitive functions (i.e. verbal report, button pressing etc.) that allow consciousness to be empirically studied at all. A proper neurobiological theory of consciousness must utilise these functions in order to accurately identify which particular neural activations correlate with conscious awareness" (ibid.: 358). Rather than negating that there *is* a hard problem, I take it as an emphasis of what makes the hard problem so hard: finding some way of empirically operationalising qualitative aspects of mental states.

[80] Sullivan criticises this notion of mechanistic explanation as not doing justice to what is actually done in neuroscience, and that it is in fact "little more than an optimistic promissory note" (Sullivan 2009: 528; see also her 2015b). If it is true that methodological and explanatory pluralism is the proper way to describe neuroscientific practice (Sullivan 2009: 536) then we certainly have a problem in reconciling it with the widely accepted view that explanation depends on a certain unity of the explanatory scheme. Specifically, to say that an effect is explained if it can only be explained with *this* single method in *this* single lab flies in the face of such expectations. It is an open question what to do if actual neuroscientific research does not live up to this explanatory ideal. But I wouldn't be so quick as to suggest that it is the ideal which should be so modified as to fit the research.

exploited by later processes" (ibid.). This problem is especially interesting, since information is the primary mechanistic currency in cognitive science. Mechanisms are categorised as cognitive because they process information – that is, because both their input and output are describable as information, and specifically as the information whose processing is the primary function of the respective mechanism –, and if we are to give a physicalist account of mechanisms, then we should at the same time give an account of information which ties into it. Yet, once we look at this particular easy problem more closely, as I'm about to do, it appears to have a hard part.

II.5.2 Analysing Cognition

Beyond describing *what* is done on a functional level, analysing cognition requires describing *how* it's done (on a mechanistic level, or the level of physical implementation) and *why* it's done (on an organismic and/or evolutionary level). For an example, let's look at he ability to distinguish nourishment from poison. Here, the relevant function is described as assigning the proper output value (nourishing or poisonous) to a given input value (the perception of something that could be either nourishing or poisonous). For example, by pointing that, say, if red mushrooms with white dots are typically poisonous and it is (ceteris paribus) advantageous to have a cognitive mechanism which assigns the value "poisonous" to perceived red mushrooms with white dots, then we already have rough ideas about the organismic purpose as well as the functional description of this cognitive ability at our disposal. As we can see, these functional and organismic analyses are closely intertwined (cf. Sullivan forthcoming); in fact, the main reason why we speak of the assignment of a semantic value in the first place is because it encapsulates the teleological notion of an organismic purpose (cf. Dretske 1986, Millikan 1989 & 1993, Neander 1995; see also section I.8.4). Said purpose is stated in terms of a relation to an external object, where the external object (plus explanatorily relevant contextual conditions) are stated as the input and the meaningful reaction to it as the output. In our example, "there are mushrooms of such-and-such a kind in the vicinity" serves as an input, whereas the output "poisonous" can serve as shorthand for all sorts of cognitive mechanisms associated with producing averse behaviour, as well as knowledge about counterfactual conditions ("if I or someone like me were to eat it, they would be likely to fall ill", etc.).

While an organism's sensory organs only ever enables it to perceive certain cues which are usually caused by the relevant object, such as a certain shape, redness and white dots, these cues are not necessarily emitted by the object which a cognitive mechanism is meaningfully/teleologically tuned to, nor are they usually exclusively emitted by it. For example, poisonous mushrooms may not appear red when perceived under unfavourable lighting conditions or they may have been damaged so as to have lost the parts which are red while retaining the parts that are poisonous; and not all red mushrooms with white dots need be poisonous. So, analyses which refer to perceived properties only will suffer from an explanatory shortage, since they cannot tie an organism's reaction to the property of being poisonous, but only to that of appearing red (etc.; and that perceptions of redness are linked to poisonous mushrooms is itself not a perceived property, of course). Under such perceptual descriptions, functional description is eschewed, and meaning/teleology must remain mysterious. It is invoking semantic descriptions which enables us to lift the veil and include information about functional purpose (cf. Cummins 1991: 10). In other words, we must first know the relevant functional relations before we can carve up neurobiological descriptions mechanistically: "The way to discover symbols in the brain is first to map out rational relations among states of mind, and then identify aspects of these states that can be thought of as symbolic in virtue of their functions" (Block 1995b: 398). Crucially, we can distinguish between these two types of description in the following way: Under the functional description, error is possible: For example, an animal can eat a poisonous mushroom because said mushroom did not cause the perception which the animal's internal mechanism is sensitive to. Merely describing the internal processing related to perception, no such error appears: Since the perception which would have elicited averse reactions simply did not occur in said case, no averse reaction was caused. But taking the external object into the picture – the poisonous mushroom – we can say that the animal made a mistake (while the information present in the non-teleological causal description of the internal processing is retained). Similarly, we are likely to find neural mechanisms which allow us to form correct sentences with a certain statistical probability and to conceive of the remaining cases as mistakes – a conception which does not follow from any non-semantic/non-teleological physical description alone.

In order to fully understand how said cognitive ability works, we need to identify the mechanism which is responsible for what we have described as the functional assignment of semantic values to environmental cues. This mechanism processes information about the environment, namely that there are (or seem to be) red mushrooms with white dots in the

vicinity, in such a way that it outputs information about whether these objects are nourishing or poisonous. Once identified, we need to understand how, on a physical level, this processing works. And in order to complete our analysis, we need to learn the reasons for our having this mechanism: in this case, because the interaction with our environment has been regularly presenting us with the need for distinguishing between nourishment and poison. For an organism which evolved in an environment that offers nothing but nourishment, such a mechanism would be redundant, and the claim that it has said mechanism *because* it needs to distinguish between nourishment and poison would simply be wrong.

Given that the most dominant ways of investigating cognitive abilities rely on functional analyses (cf. Cummins 2000: 125 f.), an evident idea is that the general purpose of a cognitive mechanism is to process whatever informationally salient or meaningful cues are part of the input and consequently yield the proper output. And while such information might be apparent on a behavioural level – we check the fridge to assess information about what it is filled (or not filled) with, etc. –, this notion of information is hard to reconcile with the physicalist picture, which becomes more prominent as we go further down in our analysis, namely to the physical basis of the associated mechanism. For reasons stated in II.1, we must assume that the relevant mechanism in humans will be neural in nature, and neural causality does not offer teleological notions supporting the required construal of informational concepts (ibid.: 127 f.). (And while we may be faced with non-neural forms of processing in artificial systems, or we may even envision different forms of cognitive processing in animals radically different from humans, as far as we actually have them, are working on them or can envision them, these would also be based on structures obeying physicochemical laws.)

The principal problem, then, is to conceptualise information, or semantic representation in general, in physical terms. It should be clear that at the bottom, we have nothing but physical terms to base our descriptions on, since the mechanisms whose function it is to process information are physical in nature. At the top, namely the functional or behavioural level, it should be just as clear that descriptions rely on notions of semantics and information. So, how do we bake this semantic cake "using only physical yeast and flour" (Dretske 1981: xi)? To say that an organism's cognitive function is to distinguish nourishment from poison is just to say that the organism can extract semantic information from environmental cues in order to yield semantic outputs such as "this is nourishing" or "this is poisonous". At intermediary levels, the notion of information persists when we individuate physical mechanism by the information they are dedicated to processing. Yet, whatever it is that is processed by the physical mechanism, our analysis cannot end at saying that something

physical "represents" the environment, since "representing" is not a physical property. This problem is the root of attempts to "naturalise representation", which means the attempts to replace the non-physicalist notion of representation with a physicalist one, and it is the root of a complication of what Chalmers thought of as an easy problem.

II.5.3. Syntax vs Semantics, or: Why is the easy problem so hard?

Since functionalist analyses of the mind systematically neglect the experiential quality of mental states (see II.5.1), modern versions of phenomenology (cf. Gallagher 2003 & 2012), of tackling questions of the self (cf. Qin, Duncan & Northoff 2013) and of consciousness (cf. Overgaard 2015, Nagel 2012) have emerged to fill this gap in recent cognitive science. While I am not going to delve into these, I wish to point out that, in light of Chalmer's classic distinction between "easy" functional analyses and "hard" phenomenological analyses, what has not been duly recognised is that both problems face some analogous difficulties. Thus, the severity of some problems thought of as "easy" has been underestimated. In the case of consciousness, we can find research to proceed by employing strategies to analyse mechanisms that underlie conscious states (again, cf. Gallagher 2012 and Qin, Duncan & Northoff 2013). Yet, Chalmer's stance was that even if we find such mechanisms, we can hardly treat the problem as solved. For the question always remains why the characteristics of the specified mechanism produce, or coincide with, a specific qualitative state – and since that is an integral part of investigating consciousness, without its being solved it cannot be any "hard problem" that can be considered solved. (At best, an adjacent "easy problem" has been solved.)

However, a similar point can actually be made for functional analyses. Given a general picture of intentional states which is akin to the one I have been expounding so far (see especially sections I.5 and I.7.4, but also the previous subsection), the analysis of such states can be married to functional analysis, and the latter can be characterised in terms of information-processing. As sketched in the previous section, investigating intentional states then crucially becomes a matter of finding cognitive mechanisms which carry out the respective information processing which is believed to underlie a given intentional state. For example, it takes cognitive information processing for anyone to arrive at an intentional state (i.e. to react with the proper attitude to the intentional content, which is itself arrived at

through information processing, or to arrive at one type of intentional attitude as a consequence of information processing).

Still, a frequently slighted point is that the description of how this information-processing is carried out is purely formal (or "syntactic") in nature, while what is processed is characterised semantically. So, a question that is analogous to the hard problem of investigating consciousness can be posed: Why do the formally specified characteristics of the identified mechanism produce, or coincide, with some specific semantic content? I believe this is in fact the crucial problem in investigating intentionality, of giving a plausible account of how formal systems acquire semantic (i.e. intentional) properties. Given that the orthodox view is that semantics is irreducible to syntax, it seems like enough of a hard problem, and it is not solved at all by giving functional explanations. This problem should therefore be construed as the hardest problem connected to endeavors of naturalising representations.

What exactly are the reasons for holding said orthodox view? First off, syntax and semantics are basic properties of representational systems. *Syntax* means a set of rules describing how potentially meaningful entities are properly constructed. For example, given knowledge about the syntax of Portuguese, anyone would be put into a position to judge that "uma carta para um filho sobre seu pai" is well-formed, without necessarily understanding what it means. We might even be able to judge that we're likely dealing with a description (or a general term) rather than a complete sentence. Once we are provided with the information that the supposed description translates to "a letter to a son about his father", we are provided with its meaning. Given this limited information about syntax and semantics, it should already be evident that there is a notable distinction between both: In order for something to be meaningful, it needs to be grammatically well-formed, i.e. to not violate syntactic rules. However, merely being well-formed does not yet imply being meaningful, as Chomsky's famous example of the well-formed sentence "colourless green ideas sleep furiously" proves (Chomsky 1957: 15).

One prevalent idea in cognitive science is that brains are effectively "syntactical engines": that they process information by way of formal steps (see I.4.4 and II.4). "We treat the mind as a semantic engine, yet when we look at the brain all we see is a syntactic engine, where the shape and orthography of neurons and neurochemicals are intrinsically causal, and it's hard to see how to get semantics out of syntax" (Griffin & Baron-Cohen 2002: 104). Here, it is important to understand that the way the brain is described, namely in terms of the interaction of the "shape and orthography of neurons and neurochemicals", only allows for any processing that happens in the brain to be implemented by means of physicochemical

causality. Yet, we also take the brain to implement semantic functions which we describe in terms of processing. So, the corresponding requirement is this: For any processing that yields B from A to be implemented in the brain, the physical implementation of A must (reliably) *cause* the physical implementation of B. Under a given semantic interpretation, A and B carry information, and thus we can speak of them as having mental content. However, the immediate causal relations in the brain hold in virtue of the physical properties of its constituents, not in virtue of semantic ones.[81] That a relationship holds in virtue of its physical/causal features is here called "syntactical", since the corresponding relational rules are merely formal, insofar as they are *not* semantic/representational/intentional/ interpretative.[82]

So, much like what happens in a computer, processing that happens in terms of neural mechanisms is seen as the sequential computing of rules such as "if input is A, then output is B". As emphasised, neurobiologically described structures are only sensitive to the physicochemical features of the input – so any of its semantic properties only apply to neural descriptions in virtue of being physicochemically implemented. And just as formal logic can be described as a set of syntactic relations between symbols, so neural processing can be described as a sequence of formal rules which are implemented by such causal processes: For example, "if B follows from A, and C follows from B, then C follows from A" can be a rule of logic as well as a description of a causal sequence. Either way, it is purely formal and syntactic. For this form of deduction or processing, the semantic properties (or "interpretations") of A, B and C do not matter, since *any* meaning that can be attached to [A→B] & [B→C] will adhere to the formal conclusion that [A→C] (assuming classical logic, at least).

So, while under a functional description the output is a semantic value, the mapping of output to input happens according to a purely formal procedure. For example, maybe our

[81] I say "immediate" to avoid confusion with an earlier point I made regarding Cummins' theory of psychological explanation in section II.3. There, the point was that certain neural mechanisms only exist because they perform a required cognitive processing. Under such circumstances, it is also correct to say that the causal relations underlying such processing hold because of semantic properties, namely those which caused someone to learn the required processing. Still, these semantic properties are not the immediate cause of any causal goings-on in our brain; the immediate causes are still physicochemical properties. Rather, under the influence of external semantic properties our cognitive "hardware" can be shaped so as to run specific kinds of "software", such as speaking English, doing calculus or rating movies.

[82] The view that the brain executes mental functions "formally" or "syntactically" does not imply a commitment to any particular form of logic such as "the assumption that [the logic which formalises common-sense inferences] is about deductive inference and completeness proofs" (Labuschagne & Heidema 2005: 146). The brain can represent things by way of a different form of logic than that inherent to the things which are represented, just as a calculator can be implemented in many ways which are themselves not representable by a calculator. Analogously, the fact that someone can dabble in first-order logic does not imply that the process by which her brain enables her to do so can or needs to be fully described in terms of first-order logic.

cognitive mechanism for judging whether something is edible or not has a rule such as "if [red mushroom with white dots] is perceived, then output [is poisonous]", where "[red mushroom with white dots]" is handled much like the previously quoted Portuguese sentence would be handled by anyone who does not speak Portuguese: namely, as a purely syntactically defined entity. Much as the Portuguese sentence to speakers unfamiliar with Portuguese, the mechanism could handle it as a semantically opaque variable: its meaning would not figure into the processing, but only its formal properties. These formal properties would be defined by the causal features relevant for our nervous system, such as those picked up by our sensory organs. Again, the point is not that our brain aren't sensitive to semantic properties (see footnote 81), but that neural processing is, by mere virtue of its being purely physical/causal, a non-interpretative, non-representational, non-intentional, non-semantic process.

II.5.4. Reconciling Semantic Properties with Naturalism

Imagine you were handed a piece of paper that says "uma carta para um filho sobre seu pai", and someone told you that everytime you receive a piece of paper with these markings, you give it to Zachary. While you would thereby execute the function "if [paper is marked with "uma carta para um filho sobre seu pai"] then [give it to Zachary]", you need not understand how the consequent *semantically* follows from the antecedent – even though it does, as a speaker of Portuguese who has seen the documentary "Dear Zachary" would assure you. So, we can see how semantic functions can be executed purely syntactically. A case similar to this was influentially discussed by Searle in his "Chinese Room" thought experiment (see Searle 1980). His point was to show that, since syntax does not determine semantics, the latter can't be reduced to the former:

> "Computation is defined purely formally or syntactically, whereas minds have actual mental or semantic contents, and we cannot get from syntactical to the semantic just by having the syntactical operations and nothing else. To put this point slightly more technically, the notion "same implemented program" defines an equivalence class that is specified independently of any specific physical realisation. But such a specification necessarily leaves out the biologically specific powers of the brain to cause cognitive processes" (Searle 2010: 17).

Searle's thought experiment remains controversial to this day, so I'd rather not rely on it too much. (Specifically, I won't even begin to mention any of his criteria for what counts as intentionality or consciousness.) However, any notable controversy is not about the fact that the brain executes semantic functions formally, but about whether mental content can actually be reduced to what the brain does: "Formal symbols by themselves can never be enough for mental contents, because the symbols, by definition, have no meaning (or interpretation, or semantics) except insofar as someone outside the system gives it to them" (Searle 1989: 45). The opposing view claims that semantics *can* in fact be reduced to syntax.

I suggest that the entire controversy is traceable back to an overstatement of the respective irreducibility: Of course such irreducibility does not imply a kind of metaphysical dualism, in the sense that something supernatural has to swoop down from the skies and lend quasi-spiritual meaning to some physical properties. Rather, what irreducibility affirms is two things: (1) That assigning mental content to a cognitive mechanism requires the notion of function, which is non-physical:

> "An example may help to clarify how functional talk is not [merely] causal talk. A physicist would say that heating a gas causes it to expand, and could provide laws that would make this predication. A biologist would say that heating a mammal causes it to sweat, and that the *function* of sweating is to keep the animal's temperature constant. The physicist would never say that the function of the gas expanding was to keep its temperature constant, even though that is precisely what happens. Thus, functions are effects, not causes, and can not be seen from the physical stance alone. A claim such as 'the heart pumps *in order to* circulate the blood' is teleological, not causal, because effects do not bring about their causes" (Griffin & Baron-Cohen 2002: 108, FN 4).[83]

What the irreducibility claim also affirms is (2) that mental content is relational, i.e. that assigning it implies a relation between their implementation, such as neural properties, and something beyond these:

> "Many in philosophy, Dennett included, subscribed to a form of externalism, according to which contentful states are seen as relational properties, and are identified by reference to entities outside the brain. Thus, if content ascriptions are extrinsically relational, then they can not refer directly to the local, causal, nexus in the brain" (ibid.: 104).

[83] In a sense, evolutionarily advantageous effects do bring about their causes, namely the mechanisms producing them.

Searle goes a bit further, adding that "[a]s far as nature is concerned intrinsically, there are no functional facts beyond causal facts. The further assignment of function is observer relative" (Searle 1995: 16). Similar reservations are expressed by Churchland, Koch and Sejnowski in an article that is considered one of the defining expressions of the project of computational neuroscience:

> "A physical system is considered a computer when its states can be taken as representing states of some other system; that is, so long as someone sees an interpretation of its states in terms of a given algorithm. Thus a central feature of this characterisation is that whether something is a computer has an interest-relative component, in the sense that it depends on whether someone has an interest in the device's abstract properties and in interpreting its states as representing states of something else. Consequently, a computer is not a natural kind in the way that, for example, an electron or a protein or a mammal is a natural kind" (Churchland et al. 1988: 48).

In a similar vein, Dennett stresses the importance of our interpretational interests for functional/intentional ascriptions:

> "It is not that we attribute (or should attribute) beliefs and desires only to things in which we find internal representations, but rather that when we discover some object for which the intentional strategy works, we endeavor to interpret some of its internal states or processes as internal representations. What makes some internal feature of a thing a representation could only be its role in regulating the behaviour of an intentional system" (Dennett 1987: 32).[84]

Just imagine how gleefully Brentano would read these passages (see I.2) – even almost a hundred years after his death, the divide between physical and intentional objects seems to be alive and well! For all these poignant descriptions of what makes a physical system representational, there are two major concerns at work here which we must deal with: Firstly, that whatever endows these systems with their representational features is "not natural" (see II.2), i.e. that it is not traceable back to natural kinds, and secondly, that it is observer-relative. Now, these points are taken to be interdependent – namely, the fact that something's being representational is observer-relative is taken to be *the reason* for its not being natural (you can find this line of reasoning in all three quotes: Searle's, Churchland/Koch/Sejnowski's and Dennett's) –, but I would still like to address them separately.

[84] I am grateful to Michael Zehetleitner for alerting me to this and the previous quote.

To address the first concern: It is a misconceived emphasis on what's "nature intrinsically", which tempts us to eschew representational facts as unnatural. And the temptation goes: If representations are not to be found within "nature proper", then shouldn't natural science get rid of it? But that, of course, would amount to scientific euthanasia by negating the whole enterprise of cognitive neuroscience, when we do have compelling reasons for not pulling the plug. Playing into this temptation is the fact that the concept "natural kind" is too often used evocatively rather than literally. Used literally, it is merely meant to designate kind-terms which are independent of human practice (as opposed to Hacking-type interactivity, see I.6.2). Being a natural kind means: Being a proper term about which laws of nature can be formulated (again, see I.6.2; for an overview of the debate see Carroll 2004). But rather than to talk about natural (or unnatural) kinds, I urge that we substitute the word "natural" in "natural kind" for the form of law in which it is used: For example, if something is a kind-term in physical laws, then it should properly be called a "physical kind". If something is a kind in biological laws, then it should be called a "biological kind", and so forth. Distinguishing kinds by their scientific fields rather than by how "natural" they are does not imply that there can be no (perhaps reductive) interrelations between physical, biological and other scientific kinds, but it lessens the temptation to see some of these kinds as the one, true, essential form of thing and others as bogus (compare footnotes 64 and 106). This is another move in my attempt to move away from painfully abstract discussions about how "real" an object is to a pragmatic discussion of their explanatory value (see I.5). Because my ultimate point is not that organisms are "as real as" H_2O, but rather that the reality of either thing, which is referred to by a kind-term, hinges on the explanatory value it has in its respective discipline. For example, the fact that one of the major prerequisites of cognitive neuroscience is that there are such things as organisms should not by itself discredit the field, just because organisms aren't a proper part of our physical vocabulary. What really matters, at least for the current state of science, is that they're part of our basic vocabulary when doing cognitive neuroscience – just as functions are. Since biology and neuroscience count as natural sciences, then content has been naturalised if it is properly explicated in terms of the two respective fields, even if it has not or cannot be properly explicated in strictly physical terms (compare II.2).

Secondly, the fact that a certain phenomenon is observer-relative does not by itself place it beyond scientific inquiry. In fact, in our case the reverse is true: that we can construe the execution of organismic functions as being at the basis of mental content means that the analysis of these functions in relation to the organism is the proper method of a

neuroscientific inquiry into intentionality. So, it is not that observer-relativity drags cognitive neuroscience down into some murky field of subjectivity, but rather that cognitive neuroscience gives us a promising stab at dealing scientifically with certain aspects of subjectivity. To be sure, this inquiry cannot be a purely neuro*biological* one (in Gold's & Stoljar's sense; see section II.2), since "individual neurons, neuronal ensembles, and neuronal structures" will not provide us with the necessary inventory of organismic concepts to tackle questions pertaining to mental content (but this is just my previous point restated: "function" is not a physical term, but still a scientific one).

Picture what descriptions of brain states, which could potentially instantiate neuroscientific laws, take into account: the spatial distribution of neurons, or their network properties; the physicochemical properties of brain cells and of the media they operate in (such as charged fluids etc.); electrical properties and timescales of discharges; innervations of muscles and sensory organs and their properties, and so on. Any such description would be lawlike in virtue of physical laws, since all of these properties can be construed as kinds in physics, and thus, physics fully describes their causal interaction. Yet, their representational or functional roles do not directly follow from these physical descriptions, since no law in physics exists which marks the state of a physical structure as representational or functional. Rather, some states, or segments of such states, are representational in virtue of a further relational property, namely a certain functional role it inhabits. For example, to say that a certain herb is a cure is to say that its physical properties function in a certain way when related to other properties (namely those of a person exhibiting symptoms of illness). While it certainly functions as a cure *in virtue* of its physical properties, its being a cure is not a description merely of its physical properties, but of its effects in relation to what it is effective on. Similarly, we should think of a physical state as being representational when it satisfies a certain relational property. So, representational and physical kind terms need not refer to different objects, but rather, some physical kinds will be representations in virtue of their fulfilling specific functional roles, just as some are also organisms in virtue of the relations they have to one another within the organismic system, and to what is conceived as laying outside the organismic boundaries (see I.8.2).

So, in order to mark these functional rules, additional methods will have to be invoked, such as those from biology and psychology, and all we have to make sure is that they will provide us with analyses of said "observer-relativity" which does not equal subjectivity with random, non-repeatable results. And this can be done because the relevant notion of observer-relativity does not imply some form of radical subjectivism at all, but, just as the

term says, it expresses the fact that some facts about organisms are relative to the organism itself. So, we might as well replace the intimidating term "observer-relativity" with "organism-relativity" – and who would be intimidated by the fact that some properties of a given organism are specific to, say, its evolutionary history? For example, we should expect that the ultimate reasons for a toad's cognitively representing worms as prey will be evolutionary in nature. Ultimately, such a sparse understanding of "observer-relativity" is all that is needed for an analysis of cognitive functions, and that is what I will proceed to show.

So, representational features need to be cashed out in objective terms. And to briefly apply this criterion to our three quotes: In Churchland et al.'s case, there should be objective conditions for when a computer can count as representing certain states. In Dennett's case, there should be objective conditions for when the intentional stance is beneficial. In Searle's case, there have to be objective facts about functions. These objective facts will most notably be derived from biology and psychology, since biology provides us with analyses for what a representation's "role in regulating the behaviour of an intentional system" (Dennett 1987: 32) consists in, and psychology provides us with accounts of mental function and performance.

To sum up, semantic irreducibility as understood in said two claims – that mental content requires the notion of function and that it is relational – hardly qualifies as a stumbling block for the neurobiological investigation of cognition. On the contrary, it is very much reconcilable with its current scientific/naturalistic framework: Namely, that relating intentional states to cognitive mechanisms implies individuating the latter in reference to the intentional objects, crucially using the notion of function. And since we can assume that the brain supplies the physical basis for executing cognitive functions, and that it can do so formally, there is quite simply *no further need* for somehow reducing semantic functions. And so, Chalmers was right (see section II.5): all we really need to do is analyse how exactly the brain executes these functions, and Searle's point, that the physical makeup of a system alone does not determine its function, already follows from a straightforward biological reading of "system" in terms of its beneficial environment-relations. For instance, describing an organism as erring, as being tricked or as misrepresenting its environment is only possible when we can describe the aim of the cognitive mechanism it employs as distinct from the physicochemical processes underlying it. In the following section I will invoke a specific example in order to show how teleological and neurobiological descriptions of cognitive mechanisms go hand in hand and how matters of subject-relativity play into these.

II.6. The Neural Basis of Cognition

One important cognitive ability consists in being able to distinguish between things in our environment which are nourishing and those which are not (compare II.5.2). For example, take the common toad (*Bufo bufo*), which preys on worms. Ewert et al. identified a neural mechanism which allows it to direct its predatorial behaviour at worms with an evolutionarily advantageous statistical rate of success (Ewert et al. 1996). They have found tectal neurons which are active when the toad is presented with certain worm-like features, such as those exhibited by elongate objects moving parallelly to their longitudinal axis (see Figure 6). Crucially, it has been found that the toad fails to direct its predatorial behaviour specifically at such objects once said neurons are removed.

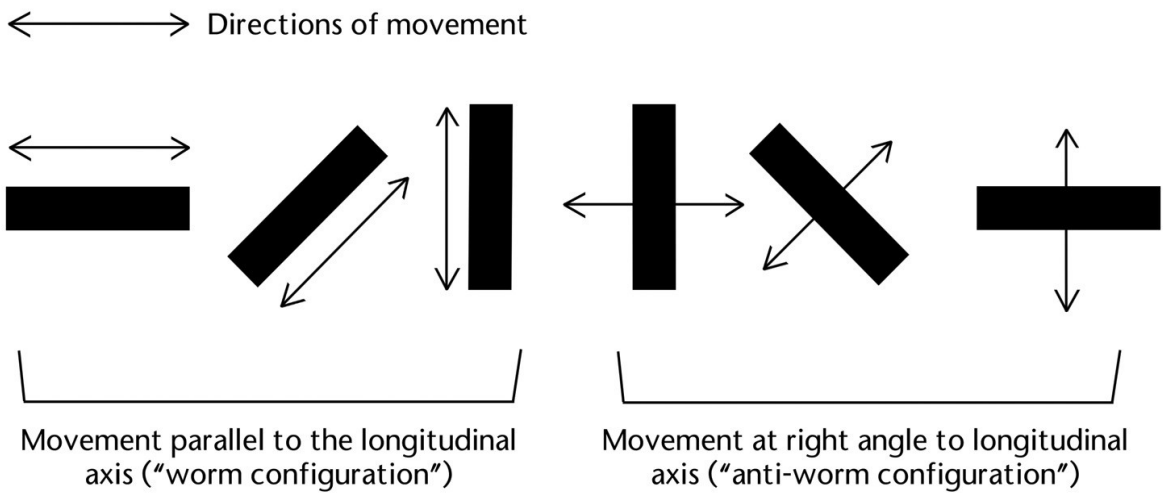

Figure 6: The toad's predatorial behaviour is elicited by the "worm-configuration" (after Ewert 1970 and Wachowski & Ebert 1996).

Ewert et al. note: "One might be tempted to call the tectal T5.2 neurons 'worm detectors'. However, one should be also aware that these neurons, like toad's prey-capture, are responsive preferably but *not exclusively* to wormlike moving objects. (…) [T]heir output is a measure of the probability that a visual stimulus fits the figural prey category determined behaviourally. The term *feature detector* is appropriate".[85] Since this mechanism is only sensitive to certain observable features strongly correlating with those of worms, it is not necessarily specific to the perception of worms. That is, toads can be tricked into preying on all objects which share the features their mechanism is sensitive to, but which are not worms.

[85] Quoted from Ewert's website, retrieved on July 6th 2014 [http://www.joerg-peter-ewert.de/3.html].

Evolutionarily, it is advantageous for toads to have developed a mechanism which is especially sensitive to these features, as long as it inhabits an environment which provides them with a big enough statistical rate of success at catching prey using this mechanism (i.e. an environment in which the toad is not regularly tricked). So, the mechanism's effectiveness depends on the fact that a toad's typical environment does not provide it with a whole lot of elongate objects moving parallelly to their longitudinal axis whose ingestion would be detrimental to the toad's health.

In line with our previously developed picture of analysing cognitive processing, we can say that, on a functional level, the toad's neural mechanism processes observable features of its environment to yield the semantic value "prey" or "non-prey". And the empirical work just outlined specifies what the corresponding neural implementation consists in. Note that a *behavioural* description is the basis for singling out the prey-capture mechanism as distinctively being part of the toad's cognitive repertoire. However, strictly speaking, behavioural descriptions themselves are not part of the neural description (in the sense specified by Gold & Stoljar, see II.2). In the case of the toad's prey-capture mechanism, the latter firstly consists in a description of how exactly the toad's sensory organs are sensitive to objects exhibiting properties of the "worm-configuration". It secondly consists in a description of how the activity of the tectal neurons elicit said behaviour (by innervation of the toad's muscles, the release of hormones, etc.). And thirdly, these descriptions highlight neural properties in whose virtue the respective neural processes instantiate physicochemical laws: how the physical basis of said sensory excitation causes a selective excitation in tectal neurons, how the activity of tectal neurons causes activity in the toad's muscles, and how additional organismic properties contribute to this process. All of this amounts to a physical description of how the toad's nervous system makes it possible for the perception of a certain stimulus to elicit directed behaviour, depending on what we can semantically describe as an internal cognitive architecture.

As previously pointed out, the upshot of this neurobiological description in a strict sense is that it does not by itself entail the semantic or functional description. So far, what we ideally get is a lawlike entailment of an effect, namely directed behaviour, from a cause, namely perception of the worm-configuration.[86] Compare my earlier example of tipping over dominos: Nothing about a purely causal description of a sequence of falling dominos implies that anything specific to this causal reaction has any representational properties (cf. Ramsey 2007: 118–150). Certainly, nothing purely physical about a domino's falling over establishes

[86] Note that this follows only if certain criteria regarding an explanatory method have been met, see footnote 80.

that it represents the pushing (see II.4). Imagine that the causal chain between the perception of a worm-configuration and the resulting behaviour were to resemble such a sequential tipping of dominos – under which condition would such a causal mechanism represent the environmental property which is associated with the cause (i.e. worm or prey)? The causal chain is describable as implementing a function, and thus semantically interpretable, once it is assigned an aim: if the causal chain makes up a mechanism whose purpose is being sensitive to the object which reliably starts the causal chain. The assignment of an aim is what lends intentionality to the neural mechanism: The neural mechanism does not only have the neural properties associated with it as outlined above, but there are also facts about its evolutionary origin or its acquisition by way of learning which warrants this teleological description (see footnote 50). So, saying that an individual mentally represents something, that their neural mechanism processes information about something, and that its output assigns a semantic value depends on using this additional teleological information about its origin, which is not inherent in purely neurobiological analyses of individual nervous systems, neural events or processes.

In the toad's case, what exactly is it that makes its neural processing a processing of "worm-like features", of shapes and movement, and not just a domino-like causal chain of sequential patterns of excitation? It's the fact that these patterns of excitation stand in a functional relation to said features of the environment (i.e. to the worm itself, as opposed to just standing in a causal relation to the perception of the worm-configuration). What makes the toad's elicited behaviour predatorial, or aimed at prey? Its functional relation to actual prey. Thus, saying that the toad cognitively represents prey is to say that the toad's neural mechanism stands in a specific functional relationship to the toad's environment: Namely, that in order to maintain its physical integrity in the face of constant energy-loss, or entropy, it needs to consume energy, and that there are organismic features of the toad which allow us to conceive of some of its physical processes as mechanisms achieving this organismic end (compare Schrödinger 1992: 67-75). In our example, the neural process outlined above is construed is a characteristic part of the toad as an organism, namely as satisfying its need to single out objects to devour.

One of the problems laid out earlier was that a set of signals or a set of syntactic rules alone cannot allow us to decide about the information carried by the signals, or about the semantics associated with the syntactic rules. Much the same applies here: A complete and lawlike physical description of the nervous system explains why a certain input-state (in the toad's case: the excitatory pattern at its sensory organs) causes a certain output-state (namely,

a certain movement of its muscles). That this causal relation is part of a mechanism which is beneficial to the organism is a wider description, adding features of the environment (namely, information about the object which caused the initial excitatory pattern) and a teleofunctional principle (such as Millikan's or Neander's, see section I.8.4). In a strict sense, it does not follow from a neurobiological description of an individual alone. Thus, it is what is called "non-individualistic" in the philosophy of mind (see Burge 1979, 1986 and Rechenauer 1997 and section I.8.5).

However, the individualistic description, i.e. that which only takes an organism's intrinsic facts into account (such as physical descriptions of neural mechanisms), also contributes to the representational description. Recall that Ewert et al. noted that the toad's tectal neurons are not worm-detectors but worm-like-feature-detectors. So, knowledge about the neurons' sensitivity *constrains* the possible reference of the semantic description, but it does not *determine* it. If the toad cognitively represents anything, it represents something that usually takes the shape of elongate objects moving in a parallel direction to their longitudinal axis, but not all of these things. Otherwise, all things which we can trick the toad with, i.e. which satisfy the worm-configuration's criteria while not being nourishing, would then count as being represented by the toad's mechanism as well. So, while it is true that the toad's tectal neurons are sensitive to things which are not worms, saying that the toad neurally represents worms is still true, since this fact is included in the higher-order functional description, which takes into account that (1) the neurons's sensitivity, combined with the toad's environment, lets it catch prey efficiently, and that (2) the fact that these neurons are sensitive to elongate objects moving in a parallel direction to their longitudinal axis is due to the evolutionarily relevant continued presence of actual prey in the toad's environment. Again, this description is not neurobiological in Gold & Stoljar's sense, but it is neuroscientific in a wider sense, insofar as the characteristic features of this neural mechanism are ontogenetically explainable only by adding functionally relevant information about the toad's environment and evolutionary history. And insofar as this is an essential part of the analysis of a cognitive ability, it is part of the description that the toad cognitively represents prey, even if it is not derivable from physical information about the nervous system alone.

To add a bit of speculation: Would this be different if the toad's neurons were in fact sensitive only to worms, instead of elongate objects moving in a parallel direction to their longitudinal axis? That is, could the proximal (i.e. the perceived stimulus) and distal (i.e. proper functional) contents (see I.8.4) conflate if the neural mechanism would be so sensitive as to constrain its potential objects down to its one correct intentional object? Well, for one,

for any complete set of observable worm-features, we know that a mechanism that is sensitive exclusively to this set would be evolutionarily efficient only in an environment in which any mechanism that is more sparse (i.e. sensitive to a subset) would lead to a significant evolutionary disadvantage. Secondly, for a neural mechanism to be sensitive only to Xs, those properties characteristic of and specific to X would have to be directly observable, since the very notion of a cognitive mechanism only makes sense if there is something observable for the mechanism to process as an input. So, we can still envision tricking any such mechanism with things which share all characteristic observable features with worms, but which are not worms (i.e. pseudo-worms which have an added non-observable toxic ingredient). As Zehetleitner & Schönbrodt note:

> "It seems to be rather easy to find examples, where indicator representations are used (...) compared to finding cases, where the success relevant feature is directly sensed. (...) The only example we were able to think of, where the success-relevant variable is identical with the indicator feature (…), is phototaxis in photosynthetic organisms (...). Phototaxis is a "behavioral migration-response of an organism toward a change in illumination regime" (Hoff et al. 2009, p. 25). Positive phototaxis is a migration towards the light source, which is a successful action for photosynthetic organisms. It seems that apart from photosynthetic organisms, light sensors (such as eyes) rather generally produce indicator representations (similar to sound waves picked up by ears, or odours picked up by olfactory sensors)" (Zehetleitner & Schönbrodt 2013: 213).

So, whenever the perceived indicator for a characteristic property a cognitive mechanism is attuned to does not conflate with the characteristic property itself, deceit is possible. Since this is rarely the case, mechanisms which are so sensitive as to only be directed toward their (proximal) intentional object will be just as rare.

On the other hand, it can be the case that environments are simply too poor to afford any object a given mechanism can be deceived with, such as when, for example, there happen to be no elongate objects moving in a parallel direction to their longitudinal axis other than worms in a given toad's habitat.[87] In such cases, the mechanism would in fact be directed to its intentional object in each possible instance, even though its distal and proximal contents do not conflate. In any case, the crucial point here is that the teleological description of a cognitive mechanisms is not decisively influenced by how many observable characteristics it is sensitive to (i.e. how well it captures the objective features of the object it is aimed at), but

[87] Vice versa, none of this implies that representing an object entails being absolutely undeceivable about it. In fact, it does not even entail knowing about all its characteristics (again, compare Burge 2007: Introduction).

rather how this sensitivity allows the organism to interact with the represented objects in a way that is functional for the organism. Thus, it may even "objectively" misrepresent indicators as long as this misrepresentation is functional (cf. Zehetleitner & Schönbrodt 2013).

The points I've made so far are only consistent with the previous quotes about intentionality not being natural (see II.5) if we modify these a bit. Compare Searle's stating that "as far as nature is concerned intrinsically, there are no functional facts beyond causal facts". In our case, this point would be specified as: there are no teleological facts as far as physics is concerned. If Searle were to be taken literally, I would have to disagree: yes, there *are* functional facts in nature, since organisms are functional entities, and since what is functional can in fact be grasped by the causal relationships as described by natural sciences such as biology. But, again, that is just to say that the natural sciences are not exhausted by physics.

In the following sections I will review the currently most important and promising strategies to supply functional descriptions of cognitive skills: evolutionary theory, dynamic/living systems theory and the social learning approach. I believe that either of them or their combined application guides most forms of content-attribution in cognitive science.

II.7. Methods for Determining Cognitive Content

II.7.1. Evolutionary Psychology

II.7.1.1. Explanations in Evolutionary Psychology

In evolutionary psychology (from now on referred to as "EP"), the central focus lies on explaining a set of psychological features exhibited by a given organism by tracing these back to a relation between the respective species and their environment: a relation called "adaptiveness", which tracks the impact this property has on the longevity of the species' genes in a specific environment. In what follows, I will focus on some necessary preconditions characterising theories belonging to this field. Underlying this focus is an understanding of EP as "a field of inquiry, defined not by any specific theories about human psychology, but only by a commitment to developing such theories within the framework of evolutionary biology" (Buller 2006: 197), as opposed to EP as a paradigm entailing "a

number of specific doctrines regarding the nature and evolution of the human mind (…) [, consisting of] a tightly interwoven web of theoretical claims, methodological commitments, and empirical results" (ibid.).[88] My reason for this commitment is that taking EP as the latter would entail that a theory could be taken as belonging to EP in virtue of mere consistency with one or several of these theoretical claims, such as the theory that the mind consists of a large number of genetically specified modules (cf. ibid.: 199 ff.).

For instance, Mercier and Sperber argue that we have erroneously conceived of reasoning as serving an epistemic function when it actually serves an argumentative one: "In a classical framework, where reasoning is seen as geared to achieving epistemic benefits, the fact that it may be used to justify an opinion already held is hard to explain" (Mercier & Sperber 2011: 66). They offer some testable predictions such as that since, according to them, "the main function of reasoning is to produce arguments to convince others rather than to find the best decision (…) we predict that reasoning will drive people towards decisions for which they can argue – decisions that they can justify – even if these decisions are not optimal" (ibid.: 61). However, their evolutionary claim – whether the respective function (or rather, the physiological basis for the associated behaviour) was acquired evolutionarily – is in fact tested by none of their predictions and thus cannot be supported by any of the evidence they deliver (ibid.: 61–71). Their investigation of the function of human reasoning positions itself as belonging to EP solely by invoking consistency with some of its theoretical claims (cf. Mercier & Sperber 2011: 58). In fact, they humbly remark that "while there can hardly be any archaeological evidence for the claim that argumentation already played an important role in early human groups, we note that anthropologists have repeatedly observed people arguing in small-scale traditional societies" (ibid.: 60). Such anthropological observations are supposed to amount to evidence by hinting at an ancient genetic foundation of the psychological trait in question. Yet, Mercier and Sperber's hypothesis that reasoning serves an argumentative rather than an epistemic function is not validated merely by providing a functional explanation for the systematic flaws that exist in our reasoning processes; rather, the explanation itself still awaits validation by empirical evidence which would support their evolutionary origin story. And this validation seems especially pressing in light of the revelation that cognitive machinery, whose rigidity had been sought to have been traceable back to the generation-spanning rigidity of amino acids, turns out to at least sometimes reflect the rigidity of social norms (cf. Henrich et al. 2010: 66). Making the distinction between a universal genetically inherited trait and a pervasive social rule requires additional evidence.

[88] For a broader picture of the content of EP as a paradigm see Buss 1995 and Cosmides, Tooby & Barkow 1992.

So, what kind of evidence actually qualifies as supporting evolutionary theories? The popular practice of spinning tales about some form of selection pressure in an ancient, evolutionarily relevant environment certainly won't do. We will all too often find ourselves in situations in which we can verify that a given organism can perform some function, yet lack evidence which tells us which, if any, of the offered evolutionary ad hoc stories is true, or whether the function even qualifies as a biological or genetic trait to begin with. These stories serve to generate hypotheses, but without additional evidence, they lack the force to one-up other "just so" stories. As Gould and Lewontin point out, "the criteria of acceptance of a[n evolutionary] story are so loose that they may pass without proper confirmation. Often, evolutionists use consistency with the data as the sole criterion and consider their work done when they concoct a plausible story" (Gould & Lewontin 1979: 588). This looseness of criteria is so pervasive that it has recently spawned a "festival of bad ad hoc hypotheses" (bahfest.com), at which presenters are asked to defend deliberately ludicrous evolutionary hypotheses in front of a live audience.

The enterprise of EP crucially depends on tying psychological traits closely to their genetic substrate. The central explanatory form of evolutionary explanation consists in claiming that a given function has either been directly relevant for the survival of a given organism's ancestor who bears that function, so much so that this ancestral function has survived in the current organism's genes, and that consequently the current function is a trait of this organism exactly because it has genetically inherited its substrate (see section I.8.4). Thus, if F is a mental function, O is our current organism which displays F, G is the gene (or a set of genes) which leads to the expression of F and A is O's ancestor, we can formulate the following hypotheses central to evolutionary explanation:

1. The primary explanation for O's having F is that O has G.[89]
2. That G was relevant for A's survival under evolutionary pressure was the cause for its originally having been passed on (while other factors not necessarily related to adaptiveness may have contributed to retaining the genome).

The initial challenge faced by evolutionary explanations of psychological functions is construing a psychological function as a trait which is tied directly to the expression of a gene. So, in order to yield a valid evolutionary explanation, it is necessary that the explained feature be the effect of gene-expression. We shouldn't understate evolutionary explanations by only

[89] Here, "primary" means both specifically and sufficiently. If G allows for several mental functions, then G does not explain F primarily.

requiring them to provide genes whose expressions are *necessary* for a certain feature. Such attempts at evolutionary explanations would provide too little specificity to actually make for a primary explanatory value: A lot of genes might be necessary to express a certain feature without actually being *specific* to the feature's expression. For instance, since my being unable to breathe would seriously impair my ability to think, genes relevant to developing lungs are necessary to perform cognitive functions, while at the same time not sufficiently explaining these. And if we're dealing with learned cognitive functions, then even those genes necessary for developing a functioning brain, while being necessary for cognitive functions in general, do not explain these learned cognitive functions as an effect.[90]

So, what makes a property evolutionarily explainable is (1) that it is genetically inherited and (2) that the reason for its being genetically inherited is that the associated function is adaptive and was subject to selection pressure. I will take any evidence that pertains to (1) and (2) as "direct evidence". Direct evidence can still range between strong and weak: ideally, showing that an identified set of genes makes the difference between exhibiting the property in question would count as exceedingly strong evidence, whereas anthropological evidence such as the one mentioned by Mercier and Sperber would be considerably weaker, even though it also pertains to (1) and (2). We will usually have to rely on comparably weak evidence, since genetic studies on humans of the sort just described – experimentally switching genes on and off and observing developmental effects – are not an option. A stronger form of evidence than anthropological anecdotes can be supplied by studies correlating the presence of specific genomes and specific traits.

Based on ideas connecting characteristics of adaptation, genetic inheritance and selection pressure specific theories in EP have been developed, such as the modularity of the mind (cf. Seok 2006). The argument for it goes: "First, our ancestors encountered a diverse array of adaptive problems, and each adaptive problem 'domain' required its own 'domain-specific' solution. Second, no single 'domain-general' psychological mechanism could have successfully solved widely different adaptive problems. Therefore, a distinct psychological mechanism evolved for each distinct adaptive problem our ancestors faced" (Buller 2006: 199). There are two options for supporting such theoretical arguments with evidence: we can either support them with direct evidence (such as findings which show that gene-expression is linked to the development of cognitive modules; see Baron-Cohen et al. 1985 & 1986 for a related, if not directly genetic, argument) or with indirect arguments boosting their theoretical

[90] Here, "learned" obviously means: Not automatically developing merely due to gene expression. This may sound circular at first, but the fact that there *are* such learned functions – driving a car, finding your way around the local mall, doing evolutionary psychology – justifies this circle.

plausibility. Any evidence which neither pertains to (1) nor (2), yet boosts theoretical coherence I will call "indirect evidence". That is, the discovery of a fact which does not amount to direct evidence for theory A, but is more smoothly integratable into A than into a competing theory B, boosts the plausibility of A. For example, finding that a given cognitive mechanism is modular indirectly supports modularity claims about similar cognitive mechanisms. Yet, it does not amount to direct evidence, since it establishes nothing about actual genetic properties. The most common form of indirect evidence comes in the form of said ad-hoc stories: Telling a story meant to establish that a certain function was adaptive in a past environment, based on which a hypothesis about its being a biological trait with a genetic substrate is inferred.

Note that my distinction between direct and indirect evidence does not run between empirical and non-empirical evidence. Indirect evidence may very well be empirical, such as in the case of finding empirical evidence for a given cognitive mechanism's being modular, which would indirectly support analogy claims. Note also that direct evidence for a claim only ever amounts to support rather than verification (cf. Popper 1959), and falsification is not a matter of falsifying singular statements or hypotheses derived from the theory, but rather of discrediting the theory as a whole (cf. Quine 1980: ch. 2). At best, indirect evidence amounts to boosting the plausibility of an ad-hoc story; but no matter how plausible or coherent, without direct evidence it will remain a hypothesis, rather than itself providing evidential support. This is not because non-empirical plausibility is generally to be disregarded in constructing scientific theories, but because the very requirements for something to count as evidence for EP is empirical to begin with: Namely, if a gene is shown to produce a specific trait, then the theory implying that this gene produces it is directly supported empirically – such a finding would do much more than merely boosting the theory's coherence.

One proposal specifying said evidential requirements for psychological functions is provided by Buss, who holds that an evolved psychological mechanism "exists in the form it does because it (or other mechanisms that reliably produce it) solved a specific problem of individual survival or reproduction recurrently over evolutionary history" (Buss 1995: 6). Consequently, an evolutionary explanation of a specific psychological mechanism has two objectives: Explaining *why* organisms display the features they do, and *how* these features came to be this way. That is, it assumes a functional analysis of psychological mechanisms (see the previous section), and ties this function to its evolutionary origin. Any such explanation is vindicated by gathering evidence supporting the how-explanation, and by establishing that the how-explanation has a bearing on the why-explanation. So, how an

organism acquired a specific property needs to be connected to why it still has it: Generally, the how-answer will supply details regarding the mechanism's adaptiveness, and its contribution to an organism's adaptiveness will be connected to its being genetically inherited, which answers the why-question. This connection is crucial, since we are not merely looking for functions which share a common ancestry; we are looking for functions which serve a *purpose*, and we aim to root this purpose in its contribution to adaptiveness. Since the purpose is a necessary part of the mechanistic explanation, we can determine the mechanism's intentionality: its specific directedness at external circumstances.

Evolutionary explanations do not just compete among each other, but also with non-evolutionary explanations. While the latter need not hold that there is no evolutionary basis at all for the traits an organism exhibits, they have to at least establish that evolutionary facts have no primary (sufficient or specific) explanatory bearing on the function which is to be analysed. Non-evolutionary explanations deny the central claim of EP, namely that a given psychological mechanism is possessed by an organism because it posed an evolutionary advantage for its ancestors. In order to defend EP against non-evolutionary explanation it is not sufficient to merely establish that this mechanism poses or posed an advantage for an organism of the respective species, but also that it was *this* advantage which led to its being genetically passed on to its offspring. So, non-evolutionary explanations will focus on claiming that non-evolutionary reasons for its being possessed by an organism yield a specific and sufficient explanation (cf. Levy 2004: 463 ff.).

II.7.1.2. Main and Side Effects of Adaptive Mechanisms

EP can explain a given property either as a main or as a side effect of adaptive gene-expression. Consider the following hypothetical example: Imagine a planet which has neither provided light nor carbon dioxide during the time frame in which a certain species of organisms has evolved. If we were to find out that these organisms are in fact able to perform photosynthesis whenever provided with light and CO_2, the claim that any feature of the current organisms has evolved *in order to perform photosynthesis* is wrong, since the relevant selection pressure has in fact never been exerted. Yet, there should be an evolutionary account of why these organisms have developed traits which allow it to perform photosynthesis by explaining this function as resting on some *other* function which is evolutionarily

advantageous even without the presence of light and CO_2.[91] While such an evolutionary origin story explains the organism's producing photosynthesis, it does not explain it on functional grounds; therefore, there is need for some theoretical revision, namely in the connection between evolutionarily selected function and de facto performed function (i.e. the connection between "why" and "how"). Notice that this form of explanation exploits the brackets in Buss's definition of an evolved mechanism (see previous section): There are evolved traits which enable the organisms in question to "reliably produce" a mechanism performing photosynthesis; yet it does not exist due to its solving "a specific problem of individual survival or reproduction recurrently over evolutionary history", but only because "*other mechanisms that reliably produce it*" solved this problem. This means that whenever a specific psychological function is the *main factor* for its being evolutionary selected, we need to skip Buss's bracketed qualifier:

[DEF. EP-Main] If it is an evolutionary main effect, a psychological mechanism exists in the form it does because it solved a specific problem of individual survival or reproduction recurrently over evolutionary history.

But whenever the psychological function is a *side effect* of other evolutionarily selected mechanisms, then the bracket is in effect:

> [DEF. EP-Side] If it is an evolutionary side effect, a psychological mechanism exists in the form it does because (an)other mechanism(s) that reliably produce(s) it solved a specific problem of individual survival or reproduction recurrently over evolutionary history.[92]

The bottom line is this: Whenever an evolutionary explanation applies and a psychological function is perfomed as a side-effect, then evolutionary selection has no bearing on it, but selection has to have a bearing on the mechanism which produces it *as a side* effect

[91] I chose a hypothetical example for the sake of clarity. Some actual examples: Current obesity is a side-effect of an originally adaptive mechanism, namely of storing energy for times of need. The fact that we cannot willingly correct the perception of optical illusions, even if we are aware of seeing an illusion, is a side-effect of effective visual processing. Also, see Pinker 1997: 39 for pregnancy sickness as a possible side effect, 223 ff. for the optical illusion "magic eye" and 534-538 for music as evolutionary side-effects. Also: "'high-level' modular architectures, such as the cognitive structures underlying chess skill, are probably tokens of module-generating developmental processes designed for other functions" (Barrett & Kurzban 2006: 640). I am indebted to Lara Pourabdolrahim for pointing these examples out to me.

[92] I modified the bracketed part, since what is important for this definitorial revision is the distinction between main factor and side-effect, not between being produced by one and being produced by a combination of several mechanisms.

(cf. Jackson 1982: 134). Note that whether the function is produced as a main effect or as a side effect does not decide whether it is currently advantageous or not; due to changes in the environment, evolutionarily selected main effects may turn out to currently be a disadvantage (such as today's misplaced/dysfunctional ingroup-outgroup behaviour in multicultural societies), and side effects may turn out to be advantageous (such as our hypothetical photosynthesis example).

In side-effect cases, there is no practical divide between verifying that an organism is able to perform a certain function, and verifying that this function has been evolutionarily selected for – because *any* function can only be performed on the necessary basis of an evolutionarily acquired hardware, radical artificial enhancement notwithstanding. The distinction between main effects and side effects get lost in purely functional characterisations – what counts for the latter are merely input-output relations, and both main and side effects are simply parts of the same output. So, we will have to make a theoretical adjustment by introducing causal notions: We need to add that the main effect was a cause of its being evolutionarily selected (i.e. a cause for its being present in offspring), while the side-effect wasn't. That the notion of cause and effect are necessary to make this distinction between main and side effect shows that evolutionary explanation is a type of causal explanation, and consists in giving the evolutionary cause for a psychological effect. When psychological effects are main effects, they both figure as causes and effects, if they are side effects, they figure only as effects. But in both cases, they are effects of evolutionary causes.

II.7.1.3. Evolutionary Explanations as Secondary Explanations

In some attempts at tracing organismic functions back to evolutionary origins, evolutionary explanations turn out not to serve as primary explanations. Consider social functions which are not determined by genetic make-up: For example, being a voter (i.e. competently participating in elections) can be primarily analysed in terms of sociopolitical requirements. Here, the primary explanation is given by a description of the political system which provides the conditions for these requirements. An evolutionary explanation can provide further (i.e. secondary) explanations for why a person can fulfill these basic requirements for being a voter, such as by describing cognitive skills required for making a mark on a piece of paper and participating in political decisions. Yet, evolutionary explanations cannot explain the difference between a person who has such competences and is

not a voter, and one who has it and is – for example, living in a society in which people can be voters.

The same can be said for learning specific languages versus being disposed to learn *any* language: "Researchers who think that language is an adaptation do not deny that different languages are acquired in cognitive development" (Mercier & Sperber 2011: 101). Accordingly, we should distinguish between a primary evolutionary explanation and a secondary one. To explain a given property primarily is to explain it specifically and sufficiently (see II.7.1.1), to give a secondary explanation is to explain the disposition to or the necessary basis of this property. So, if the disposition to speak a language is an evolutionary adaptation, then evolutionary facts about its development would explain the current disposition primarily. However, being proficient at speaking English is explained primarily by social facts, namely the details of being part of a community of English speakers. Still, evolution potentially provides a secondary explanation by explaining how an English speaker is disposed to pick up any language (or any sufficiently similar to English). This secondary explanation would be concerned with explaining necessary psychological and behavioural dispositions to fulfill the required function (compare other examples such as mastering C++ in section I.4.4).

Such explanations are secondary because having the ability to perform (or performing) the specific function that is to be explained does not follow from them directly. What follows is rather a general template for fulfilling diverse functions which evolution has equipped human beings with. For example, much as anyone who can pick up English is (ceteris paribus) also genetically equipped to pick up French, we are not restricted to either using cars or to using bicycles. So, one single function – getting from A to B – might very well be fulfilled in several distinct ways, with distinct cognitive capacities underlying these: learning how to ride a bike does not cognitively enable you to drive a car, and vice versa. On the other hand, speaking two different languages might exploit some of the same cognitive functions (namely linguistic capacities), while the social function might differ: The function of speaking French is to get along well in France, not to get along well in Germany (and we can go so far as to construe narratives in which speaking one language systematically leads to survival, and another doesn't, thus potentially being of evolutionary relevance). So, since it is possible to have a pairing of one cognitive basis with two functions just as well as a pairing of two functions with the same cognitive basis, no immediate connection to evolution needs to be presupposed across the board when it comes to explaining the relation between mental function and the underlying cognitive ability. In such cases, domain-general learning

mechanisms can explain how functions which are eventually shaped or determined socially can be fulfilled by genetically inherited mechanisms (cf. Buller 2006: 199). On a neural level, phenotypic plasticity can implement such learning mechanisms:

> "Phenotypic plasticity is the capacity of a single genotype to produce more than one adaptive phenotype – more than one anatomical form, physiological state, or psychological mechanism – in response to environmental conditions. And research in developmental neurobiology has shown that mechanisms of neural development embody a plasticity that produces, through interaction with the local environment, brain structures that perform relatively specialised cognitive functions" (ibid.: 200).

Once again, note that it is always possible to make a minimal claim about the secondary relevance of an evolutionary explanation: Namely that the architecture underlying a given cognitive mechanism, insofar as it depends on gene expression, has an evolutionary origin. In connecting the notion of psychological properties and traits tied to evolutionary origins, EP is concerned with attempting to provide a primary explanation, which is why it sometimes competes with non-evolutionary explanations which also claim to be of primary explanatory value.

II.7.1.4. Challenges to Evolutionary Explanations

The decisive question that has to be answered in order to determine whether a given evolutionary explanation applies is whether the respective psychological explanandum constitutes a trait in a sufficiently biological sense, or rather a localised, culture- or society-dependent or learned property. Properties that are local and/or social in origin can also be adaptive, so properties of these traits can be consistent with evolutionary ad-hoc hypotheses (see II.7.1.1). However, they will differ insofar as they will not depend on an ancient point of origin, and gene-expression won't sufficiently and specifically explain them.

That a given property differs between cultures can constitute indirect evidence for its not being genetically but socially inherited/learned. Henrich et al. (2010) have reviewed such evidence, suggesting that many psychological properties, ranging from higher-order ones, such as styles of reasoning and a sense of fairness, down to those which seem more hard-wired in comparison, such as properties of perception or the heritability of the IQ, do in fact differ significantly among current populations. These differences are not accounted for

evolutionarily, since suitable predictors for them are in fact non-evolutionary in nature: "a population's degree of market integration and its participation in a world religion both independently predict higher offers [in ultimatum game trials, tracking a sense of fairness], and account for much of the variation between populations" (Henrich et al 2010: 65). There, it is also pointed out that in some cases cultural differences in cognitive processing correspond to differential brain activation during the performance of the same cognitive tasks (ibid.: 72, see also Hedden et al. 2008). Given the range of possible solutions to many theoretical or practical tasks and the amount of different strategies for arriving at any of these, it is hardly surprising that the employment of problem-solving strategies can depend on which of these are favored in a given cultural environment, and that the use of different strategies can in turn result in the employment of different cognitive mechanisms and different corresponding neural bases. Yet, if cognitive processing were determined evolutionarily, it would be more plausible to assume that corresponding brain activities doesn't differ across culture, insofar as these have evolved from the same ancestor.[93] If a function has been selected evolutionarily, then same function should imply same genetic basis, and same genetic basis should imply same brain activation. And cultural difference shouldn't be a better predictor for the brain activity underlying cognitive performance than sameness of the evolutionarily selected function.

While it hasn't been shown that these examples apply analogously to cognitive processing across the board, and we should expect there to be both sets of evolutionarily inherited as well as socially inherited cognitive functions, these considerations should at least illustrate how sociocultural explanations compete with evolutionary explanations of cognitive functions. We should also expect them to overlap, namely when social structures have an influence on which traits are adaptive and which aren't. For example, since we have reason to believe that the IQ's heritability, and the degree to which higher IQs favor survival more than lower IQs, are itself subject to cultural influence (see Henrich et al. 2010: 77), identifying the cultural factors would in this case constitute the primary explanation for genetic make-up, and the genetic make-up constitutes the primary explanation for the expression of this cognitive trait only within this social framework.

In those cases in which the presence of certain features can turn out to be primarily explained without recurring to evolution, it may still be necessary to try and acquire

[93] In this and the following sentence, "sameness" of brain activation does of course not imply *exact* sameness, but sufficient functional similarity. That is, if we were to find out that a general set of (potential) activation patterns were to underlie a cognitive function, we would expect sameness to mean being part of this set. Differential brain activation, as cited, would mean not being part of it.

evolutionary evidence – simply because there will be no other way to find out whether assuming an evolutionary origin is plausible without reviewing available evidence. So, the fact that it is strikingly plausible that acquiring evolutionary evidence is superfluous for explaining why someone is a voter should not let us forget that we are likely to come across cases in which the question whether an evolutionary origin story actually has a bearing on explaining the presence of a psychological feature cannot be settled in the absence of reviewing evolutionary evidence. Sometimes, non-evolutionary theories (such as those pertaining to social learning) will only outdo competing evolutionary theories after evolutionary evidence has been reviewed and turned out not to support establishing the respective function as an evolutionary effect. This is perhaps going to be the case in evolutionary explanations of gender attributes, where currently we cannot be sure where our evolutionary heritage ends and social conventions start (cf. Levy 2004). Thus, we may be presented with the paradoxical case that evolutionary evidence is instrumental for establishing its own expendability.

II.7.2. Dynamic Systems Theory

Dynamic systems theory typically opposes the notion of a centralised representational processor, i.e. the view that, in order to master the challenges the environment poses and to guide their behaviour accordingly, organisms use rich internal models of this environment (cf. Brooks 1991, Thelen & Smith 1994, Beer 2000).[94] As of late, a popular strategy to explain away internal models is to invoke forms of embodiment and embeddedness (compare footnote 8). That is, if functional reactions to external challenges can be explained as direct reactions to stimuli mediated by the senses there doesn't need to be any additional internal encoding, and comparatively simple algorithms would suffice.

Two things should be noted: firstly, these internal models stand in no obvious relation to the kind of representations intentional psychology invokes, and it is an open question to which degree neural representations (see II.3) can be explained away by said approaches. As I have argued, partaking in symbolic practice is necessary for being assigned "rich" notions of intentional states, and adherence to some of the laws specified by intentional psychology (which amounts to saying that taking Dennett's "intentional stance" is pragmatically justified,

[94] The examples for neuronal representations which I presented in chapter II.3 are candidates for such internal models.

see section I.7.5) is necessary for being assigned "sparse" notions. But it is an open question how informational richness of internal models contributes to either notion. All said notions of rich and sparse representation require is that some internal informational state(s) fulfill the function associated with the ascriptions. The question how informationally sparse or rich an internal model will be depends on two things, (1) on how sparse an implementation of such capacities can theoretically be and (2) how these capacities are actually implemented in an agent's cognitive substrate. And these two points have no direct bearing on matters of representation as construed above. For example, we can envision functioning thermostats to either be based on sparse or on rich internal mechanisms but outwardly "behaving" in the same way, so as to both justify taking the intentional stance in virtually the same way as well.

Which segues into my second point: There is a limit to the explanatory power of accounts which seek to do away with internal models. Such a theory faces similar problems as behaviourism, insofar as both seek to explain cognition primarily in terms of organism-environment interaction. This makes it vulnerable to some of the criticism that has been directed at behaviourism, especially the question how it can solve problems which seem to require "representation-hungry" solutions by requiring substantial internal information storage (cf. Clark 1997: 168, also see I.7.3).

However, looking beyond its criticism of internal models, dynamic systems theory aims to lay the groundwork for naturalistically analysing cognitive representations (Thelen & Smith 1994, Bechtel 1998). Michael Zehetleitner (forthcoming) has pointed why we should favour this approach over evolutionary ones, namely because the latter suffer from the *Problem of Historicity*, whereas the former does not.[95] To illustrate this point, Zehetleitner invokes Davidson's thought experiment of the "Swamp Man", a creature that is physically identical with a human being but has no evolutionary history (for the details see Davidson 2001b: 19). If we were to follow Millikan's and Neander's teleosemantic account – i.e. if representational content is that which relies on evolutionarily selected structures (see section I.8.4) –, such a creature could not be ascribed any mental content at all. Similarly, even if its behaviour would justify taking the intentional stance (i.e. assigning mental states based on laws of intentional psychology), we would be unable to identify any physical structures underlying these mental states, since these structures would have to be individuated by their

[95] An editorial note: I assume that Zehetleitner's forthcoming paper is going to serve as an ideal reference for this subsection. However, since at the time I am writing this it is still unpublished, I am basing the position presented here on some of his recent talks (see footnote 63). Hence, his eventual position might deviate from my present portrayal. To make up for this, I am also referencing several analogous (if less unified) positions here.

mental functions and determining their functions hinges on evolutionary selection. For these reasons Zehetleitner proposes an approach that is crucially ahistoric.

Cummins makes an analogous point, but rather than by citing Davidson's "swamp man" example, he invokes the "teleporter" from Star Trek,

> "a kind of duplicating machine that duplicates organisms not by cloning, or by any other biochemical process that uses the information coded in the organism's DNA, but just as a copy machine duplicates a printed page without understanding it. The machine I have in mind produces a perfect physical duplicate of an organism without 'understanding it'. (…) [T]he assumption behind the Star Trek transporter is that the duplicate is the same person who entered the transporter. There seems little doubt that, for the purposes of a psychology experiment, a molecule-by-molecule duplicate of a person would do as well as the original. To deny this seems to be to deny physicalism" (Cummins 1991: 80 f.).[96]

To be sure, since the odds that we are to encounter such an ahistoric being are neglectable, any connected problem couldn't turn into a virulent methodological problem for identifying mental representations in biological organisms. Thus, I propose reframing the problem of historicity in terms of our more pressing methodological problems: Namely, while we can assume that all organisms whom we assign mental states to also have an evolutionary history, it can be difficult to invoke this history in order to identify or characterise a specific function. I have sketched some of these problems in section II.7.1, and while they sometimes involve conceptual confusion about how secondary evolutionary explanations cannot play the role of primary ones, and indirect evidence cannot serve as decisive one, more frequently they boil down to a lack of decisive evidence. That is, while we are likely to gather general evidence about evolutionary interconnections between organisms, evidence about the evolution of specific cognitive functions is much harder to come by, which (as pointed out previously) is why evolutionary psychology remains a field of much controversial speculation. Thus, if we could find a way to characterise such functions independently of evolutionary/historic properties of organisms, and rather infer these characteristics from current properties, we would gain a significant methodological advantage (also compare Waskan 2006: 3.6).

[96] It should be noted that Cummin's framework in which this quote should be placed is crucially different from mine: Here, he seeks to find a notion of representation which satisfies computational requirements, and this is where his criticism of historicity hails from: "According to computationalist accounts, history is an accidental property of a cognitive mechanism. According to computationalism, cognitive systems are individuated by their computational properties, and these are independent of history" (ibid.: 82). While there are notable connections between the computational notion and the one I'm after, mine is at least different insofar as the satisfaction of computationalist requirements does not straight away enter my account as a premise.

This is what the dynamic systems approach aims to yield. Roughly, it individuates representations by equating them with internal structures in living systems which inhabit a specific place in a causal model from dynamic systems theory (Zehetleitner's respective approach relies on Friston 2010, Ashby 1954, Bischof & Zehetleitner 2015). The idea is that representations are things which elicit a directed biological activity that supports an organism's structural integrity and ensures its endurance as a functioning system by minimising entropy (which is one of the basic characteristics of a living organism, as historically proposed by Schrödinger 1992: 67-75). In a nutshell: What an internal structure is intentionally aimed at is learned by finding out how this structure supports the organism's homoeostasis. To invoke an example by Dretske (1986): Magnetotactic bacteria populating the northern hemisphere's oceans have an internal magnet that guides them toward the geomagnetic north. This causes them to reach deeper, oxygen-free waters, which is crucial for their survival. Since the decisive factor for the bacteria's homeostasis is the anoxic environment, being propelled toward it can be singled out as the proper function of the organism's respective mechanism and its external aim. Thus, the mechanism can be said to have representational properties (although, as Dretske points out, certainly not full-fledged beliefs) aimed at anoxic waters. (For brevity's sake I will gloss over the differences between Zehetleitner's and Dretske's account.)

We can imagine that in many cases, evolutionarily individuated functions and representations will in fact coincide with those individuated using the dynamic systems approach as just sketched. In cases analogous to the magnetotactic bacteria, the mechanism which is conducive to maintaining homoeostasis is also the mechanism which was selected evolutionarily to execute this function. However, not only can finding out about a mechanism's function and its evolutionary history be separated methodologically, we can also envision cases in which the dynamic systems approach will single out functions which are more specific than evolutionary approaches: Joining with the NSDAP was conducive to homoeostasis in Nazi Germany, while only very broadly having been evolutionarily selected for (if, say, carrying genes responsible for developing ruthless compliance was a cause for joining). Yet, many related examples will exceed even the dynamic systems approach: If we wish to explain the representational qualities of the associated ideological beliefs, both evolutionary and dynamic systems explanations would reach their limits. That is, representational systems feeding on "rich content" (see I.4.5) will incorporate elements which need neither be innate nor conducive to homoeostasis, but only depend on systematically heeding norms of symbolic systems. Zehetleitner & Schönbrodt exemplarily mention

"mathematical symbolic systems" since these are purportedly "completely unrelated to the external world" (Zehetleitner & Schönbrodt 2013: 216). Despite our not needing to acknowledge this strong claim, we can invoke their example on the grounds that adhering to mathematical rules is primarily explained in heeding rules of symbolic systems instead of being grounded in gene expression or homoeostasis. That is, mathematical knowledge can be conducive to homoeostasis, perhaps even to survival, but neither approach can primarily explain why a given individual has mastered general mathematical principles. Similarly, subscribing to certain ideological beliefs can be construed as conducive to achieving evolutionary aims and goals concerned with bodily and functional integrity; but only in such a roundabout way that it is dubitable whether said two approaches still provide anything that can justifiably be called an explanation. A social learning approach would prove more fruitful in such cases, since it is able to explain how we come to adhere to norms related to symbolic systems and why we do so, and it would expand our explanatory scope regarding intentional mental states considerably.

An approach which seeks to individuate cognitive representations in terms of homoeostasis faces another shortcoming: namely, while it can account for misrepresenting, it cannot distinguish between deliberately or purposefully meaning something and meaning something by accident. For instance, an evolutionary mechanism which is accidentally advantageous will be ascribed the same content as one which has been selected for this very advantage: main and side effects are thusly lumped together (see II.7.1.2, esp. footnote 92). So, we are offered a trade-off for solving evolutionary theory's problem of historicity: We lose some of its specificity. For example, the heart contributes to an organism's homoeostasis by supplying oxygenated blood to different bodily regions, but by doing so, it also contributes to a temperature exchange: if the torso is warmer than the legs, then the heart's pumping blood contributes to a more rapid temperature exchange between torso and legs than if it weren't pumping blood. Depending on external circumstances, this can serve homoeostasis (by, say, keeping the legs warm). However, it seems exceedingly plausible to regard pumping blood as the heart's main function (or "basic factor", cf. Cummins 1991: 76 f.) and temperature exchange as a side effect. But in order to distinguish between main and side effect, pointing to matters of homoeostasis isn't enough, since these will only track what is beneficial, no matter if it is purposefully or accidentally so. The heart, along with the other organs, also adds weight to the body and might help me not to get blown off a cliff; all the same, supplying weight surely isn't the main function of the body's organs (in fact, there could be as many situations in which weighing someone down isn't functional at all).

Amputating his own arm turned out to contribute decisively to Aron Ralston's homoeostasis when it was caught under a dislodged boulder; yet, his arm's purpose wasn't to be amputated. We can invoke the conceptual framework and methods of evolutionary theory in order to make this distinction which can't be made in terms of homoeostasis. Homoeostasis also cannot even mark processes contributing to organismic change, and perhaps most strikingly, to matters of reproduction as functional: organs contributing to reproduction cannot be assigned a homoeostatic function, and puberty and menopause would remain mysterious.

Since this distinction between purposefully and accidentally representing is integral to ascriptions of semantic content across the board, there is an analogy for this problem at the level of intentional psychological explanation: I can mispronounce Porto for Bordeaux while booking a flight, yet still end up enjoying my vacation in the wrong spot; but that doesn't change that I did in fact mean Porto and not Bordeaux. Even if a mechanism's misperformance ends up having advantageous consequences for its host (and even if our meaning something by accident can be beneficial, cf. Zehetleitner & Schönbrodt 2013), we should be able to separate meant content from beneficial content. So, I suggest that while matters of homoeostasis can certainly provide heuristics for assigning content, it can only do so because matters of proper function and matters of homoeostasis often coincide; yet, the latter cannot settle matters of proper representation all by themselves, and we will ultimately still have to defer to the respective mechanism's proper purpose.

What we should mark here are processes which systematically contribute to a species' homoeostasis, not to those of an individual under special or extraordinary circumstances; but with the need for marking the required systematicity comes the introduction of evolutionary history, and thus the problem of historicity. Given these two options, I'd rather downplay the problem of historicity than abandon the notion of adaptiveness altogether. Both in Davidson's "swamp man" example as well as in the case of the teleporter from Star Trek we can say that the reason for the duplicate's having representations is that the original had them. In that sense, it is wrong to say that such a duplicate doesn't have a history: It's just that its history took an unlikely turn by duplicating an organism with adaptive representational mechanisms. That is, we can accept duplicates; but the distinction between these duplicates accidentally meaning something and their actually meaning something can only be made by mentioning that they are duplicates of specific kinds of organisms (and therefore, Cummins' criteria hailing from computationalism should be amended, see footnote 96). Accepting this also means that an exact physical duplicate need not inherit the original's representational states; but we knew that already (see I.8).

II.7.3 Social Learning

Finding out about an organism's evolutionary history plays an important role in explaining why it has developed its cognitive mechanisms, why they work the way they do, and which parts of its neural architecture should count as fulfilling the organismic functions in question. Still, we cannot exclusively rely on evolutionary accounts to explain these, because the cognitive functions of organisms are not exclusively or specifically determined by gene expression: they can also be learned. Most animals have the capacity to learn in one way or another – to change their behaviour, their dispositions to behaviour, and/or their cognitive processing depending on their experiences and habitat. Certainly, learning plays an important part in the acquisition of cognitive capacities especially in higher animals, including humans.

Okano et al. define learning quite minimalistically as "a process of acquiring memory" and memory as "a behavioural change caused by an experience" (Okano et al. 2000: 12403). But while experience is a necessary part, it is not the only cause of behavioural changes which count as learning.[97] Rather, it is the *interaction* of organismic or genetic dispositions and experiences which results in learning. When mentioning a genetic disposition and external circumstances does not suffice to explain behaviour, learned abilities can fill this explanatory gap. For example, in many animals, reacting with aversive and fearful behaviour to snakes can be explained by pointing out their innate disposition to fear snakes and the perception of a snake.[98] Formally, this behavioural explanation works just like the syllogism employed in action explanations (see section I.6):

(1) O has a disposition to fear snakes
(2) O perceives a snake
(C) O shows fearful behaviour.

If we assume the disposition to fear snakes to be innate, then (1) will be an evolutionary fact about O, and the cognitive mechanism associated with it will be analysed just as laid out in the previous sections. To briefly recapitulate: There, I have stressed the importance of adding evolutionary accounts to analyses of cognition, since evolutionary explanations are tied to the identification of cognitive mechanisms. Mechanisms are identified by their functional role for

[97] This definition is too broad for another reason: it encompasses exercise. The difference is that learning refers to the acquisition of an ability, whereas exercise means improving on it.
[98] However, there can also be features which depend both on innate properties as well as on learning: For example, Mineka and Cook (1989) argue that monkeys have an innate disposition to acquire a fear of snakes.

the organism. Assuming a teleofunctional account, functional roles are tied to evolutionary history, and more specifically, an organism's genetic make-up. This is crucial for specifying the function that is to be fulfilled – namely, to produce the proper output.

However, if a disposition that fulfills the role of premise (1) in a given behavioural explanation is not exclusively determined by genetic expression, then we need to move beyond such accounts. For example, if the syllogism were the following

(1) P speaks English
(2) P perceives a sign that says "no parking here"
(C) P does not park her car where the sign points to

then, clearly, evolutionary accounts would not provide us with a satisfying explanation. They simply are not fine-grained enough: While there could be an evolutionary explanation for speaking a language, there cannot be one for *specifically* speaking English (rather than a different language). And (C) clearly does not follow if we substitute (1) with an evolutionarily explainable fact such as "P speaks a language" (compare Cummins 1991: 49 f.).

We can find many cognitive mechanisms in humans to not depend exclusively on genetic expression (or not to be "hard-wired"), insofar as they are either developed only if the environment provides certain stimuli, or insofar as they undergo fundamental environmentally-induced change within the individual's lifetime. To be sure, this fact does not negate or contradict evolutionary accounts. Rather, we need explanations which complement applicable evolutionary ones to compensate for their not being fine-grained enough. In these required fine-grained explanations the environment does not merely take the place of a specific context (i.e. formally taking the place of premises such as (2) in our syllogism), but in the shaping of the underlying cognitive disposition, i.e. premise (1). This is consistent with finding that what has been evolutionarily selected for in our cognitive architecture are not merely "hard-wired" components, but rather that evolution favours phenotypic plasticity altogether (see II.7.1.3 and I.8.5).

To add three remarks: Firstly, the fact that neural plasticity underlies a lot of neural processes is a given. What presently matters is that it is a way of shaping neural architecture which is not exclusively determined genetically, but also by interaction with the environment. Secondly, learned abilities are not to be confused with those which are a product of individual (ontogenetic) phenotypic development. Some abilities are not present at birth – such as the ability to procreate –, but they are not learned either. Learned abilities are those which require

environmental cues. For example, we can find that without learning to ride a bike, this ability will not magically appear by a certain age. Neither will the ability to speak English. Thirdly, this is not to say that learned abilities such as riding a bike or speaking English do not depend on certain abilities which are the product of genetic expression. They simply do not *exclusively* depend on these. That is, we can envision that someone has the genetic disposition to learn the English language, but that either due to environmental limitations or due their own decision, they don't actually learn English.

Dretske makes this connection between social learning and the import of semantically characterised states for explaining human behaviour explicit:

> "The reason learning is so central to intelligent behavior, to the behavior of people, is that learning is the process in which internal indicators are harnessed to output and thus become relevant—as representations, as reasons—to the explanation of the behavior of which they are part. It is in the learning process that information-carrying elements get a job to do because of the information they carry and hence acquire, by means of their content, a role in the explanation of behavior" (Dretske 1988: 104).

While there is thusly sufficient theoretical reason for invoking learning as underpinning some intentional capacities, there can also be empirical effects relating mental representations and learning. For example, studies have found that learned representational conventions, such as grammatical gender, influence cognitive representations: "Our findings indicate that grammatical gender can lead speakers of a language to think about inanimate objects in terms of properties that they associate with males and females. The properties that "pop out" when people think of inanimate objects are the result of a developmental process in which language plays a meaningful role, starting at the age of 7 years" (Sera et al. 2002, also see Athanasopoulos et al. 2015).

Apart from those cognitive skills which make explicit reference to symbols, such as linguistic skills (see section I.4.4), much cognitively invoked semantic content refers to objects about whose existence and characteristics we need to be educated. Not all invoked objects depend on learning: some basic objects of cognitive representations which were of direct evolutionary importance are indeed likely to be "hard-wired" (here, the objects associated with some of my earlier examples for neural representations in section II.3 might qualify). Yet, it is characteristic for human cognition to acquire mental objects which are evolutionarily "new" – which have not been around long enough to be anything but learned. Whether we think about going to the mall because we have determined the fridge to be empty,

or only vote for parties which support gender equality, or decide to wait for the extended cut of the "Lord of the Rings" before buying the DVD – none of the objects of our cognitive processing in these cases, and in many more akin to these, could have exclusively been determined by hard-wired mental representations. When we find Robert Williams expressing that he's "been fascinated since days in graduate school with underdetermination/ indeterminacy arguments in the theory of representation" we can see the same investigative aim:

> "[looking] at alternative traditions – salient among them being the causal-teleological accounts of Dretske and Millikan, or (in the case of language rather than mental content) the ideas surrounding the causal theory of reference, (...) the really compelling stuff that I could extract seemed to lack some of the virtues I prized in interpretationist accounts. Interpretationism, if it worked, would give a story about all content, not just special cases (reference to medium sized dry good and their observable properties). My hunch was that I wouldn't find in these alternative traditions a satisfactory story about unsexy but genuine questions about what grounds the relation between the word 'of' and its semantic value, or about the grounds of content of highly theoretical beliefs remote from perception or action."[99]

In a nutshell: Cases of rich content are characteristic and widespread enough for human cognition to warrant widening our accounts beyond teleofunctional accounts which determine mechanisms based on genetic expression only (compare section I.4.5).

II.7.4. A Unified Account of Cognitive Representation

Given the teleological principles reviewed in the three previous subsections, we now arrive at a unified form of cognitive representation (see Figure 7): Whether a cognitive state is representational/intentional is determined by its falling under certain teleological principles. These principles can either be stated in terms of functional aims of organismic mechanisms which have been evolutionarily selected, or in terms of norms which are acquired through social learning. The former are not necessarily associated with mental content (yet, conforming with them can still provide grounds for pragmatically assigning mental content in attenuated form, see I.7.5), although having "rich" mental content can build on those

[99] This and Williams' preceding quote were taken from an interview conducted by Lisa Bortolotti for the "Imperfect Cognitions" blog. See http://imperfectcognitions.blogspot.co.uk/2015/04/the-nature-of-representation-interview.html (retrieved on April 16th 2015).

capacities conforming with them. Acquired norms can be those of logic, rationality, or the normative aspects of laws of intentional psychology. For example, if I promise to be at a certain place, the fact that I should show up there is a matter of a conventional social norm. My being there is rational insofar as it contributes to upholding this norm (much like Kant's categorical imperative requires the maxims underlying our actions to be generalisable, cf. Kant 2011: 33 & 57 ff.). But this kind of rationality is not the kind which is a requirement for having intentional states: that is, anyone can have good reasons (i.e. potential psychological causes) to break this promise, namely when she has more urgent matters to attend to. In other words: Any reasons for breaking this promise are assigned in terms of mental states which (ceteris paribus) cause someone to break it. However, assigning any intentional psychological causes depends on a second kind of rationality: on being consistent and generally believing truths (see I.7.4). These norms are obviously distinct from norms such as that promises should be kept: Abiding by the former is required for us to be assigned *any* intentional states at all. So, apart from the kind of evolutionary aims mentioned before, rich intentional states (i.e. those for whose having evolutionary/organismic explanations cannot give sufficient conditions) can be traced to two kinds of norms: norms of semantic interpretation, which are basic and indispensable, and conventional social norms, which can explain behaviour contingent on social learning.

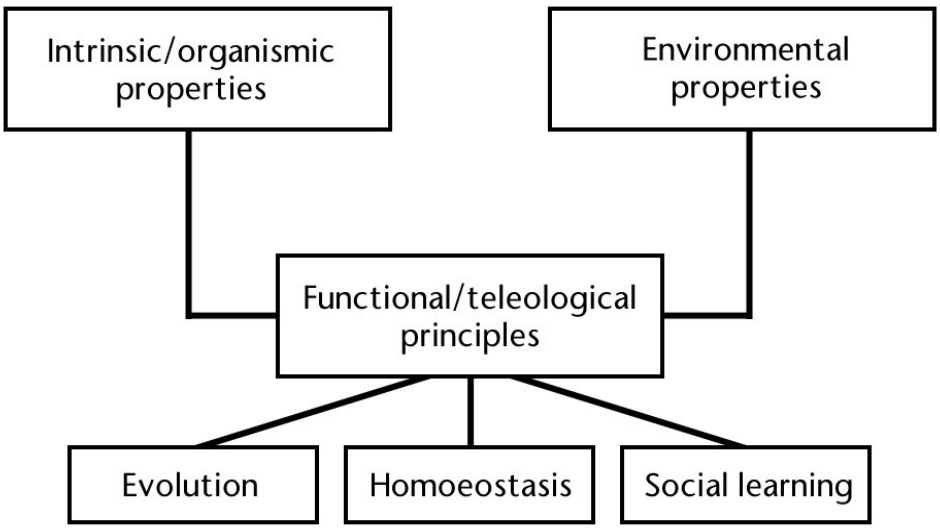

Figure 7: Representations are relational structures holding between organismic and environmental properties which are connected by teleological principles. Evolution, homoeostasis and social learning need not exhaust these principles, but they are the most dominant in current cognitive science.

II.8. The Neuroscience of Intentionality

II.8.1. From Sparse to Rich Mental Content

As Dretske (1986) and others in his wake have pointed out, one problem when specifying how biological organisms represent objects is to pinpoint the exact content and to allow for the possibility of misrepresentation (see II.4). So far I have argued that the solution to both problems is to require that the object of a cognitive mechanism be the teleological aim it has been fashioned to achieve. Millikan states this aim in terms of evolutionary selection (see I.8.4 and II.7.1), dynamic systems theory does so in terms of homoeostasis (see II.7.2), and sometimes we need to invoke norms acquired through social learning (see II.7.3). In Dretske's example the magnetotactic bacteria's characteristic mechanism, while proximally being aimed at guiding the bacteria into the direction of the earth's magnetic north pole, in fact executes the function of reaching deeper, anoxic waters. Being guided into the direction of the north pole is the means by which the mechanism executes that function. No fruitful function would be fulfilled for these bacteria by merely being directed to the north pole if this would not result in reaching anoxic waters.

Specifying a mechanism's proper function is a matter of knowing enough of its environmental and developmental context, as well as pinpointing the role it plays for the organism. Both considerations of homoeostasis as well as of evolution can play their part in this regard; but they do not always have to. Cognitive mechanisms can carry out functions which are in no apparent way related to homoeostasis and for whose acquisition evolutionary development is not sufficient. For example, we assume that those of us who have learned addition use a physically realized cognitive mechanism which has its roots in genetic expression in order to perform it; yet, evolutionary heritage alone does not suffice to put us into a position of being able to perform addition, and while some of the means by which we learn addition can be associated with homoeostasis in a broad sense (such as peer pressure or associative conditioning), characterising performing addition solely in terms of a homoeostatic function means getting the concept of addition fundamentally wrong. Rather, in cases such as these, agents shape evolutionarily acquired cognitive mechanisms in a way that goes beyond what is evolutionarily explainable in order to perform the respective function.

What holds for cases of sparse representation actually chimes with a view of rich representation that is broadly Davidsonian in nature, namely that intentional explanation is, "in a fundamental way, not reducible to physical, neurological, or even behaviouristic

concepts" (Davidson 2001a: 154). This is because "[e]vents conceived solely in terms of their physical or physiological properties cannot be judged as (...) concerned with a subject matter" (Davidson 2004: 180), or, in other words, as having what I have called rich intentional content, and, if we do not add information about the purpose of physiological structures, not even as having sparse content. When we are dealing with rich content, Davidson's quote identifies the central problem of naturalising those mental states which are used as explanatory kinds in intentional psychology. For naturalists it is a reason for wanting to explain intentionality away, and for intentional realists it is a reason to dismiss attempts at naturalising the mind. I have been arguing for the indispensable explanatory power of intentional psychology, while also claiming that it is not merely a descriptive theory, since the normative aspect of its laws – logical, conceptual or rational norms – constitute a force which is characteristically causally shaping our minds (compare I.6.4).

In cases of sparse representations, teleological descriptions have to be invoked in order to determine representational content and to allow for misrepresentation, and for this reason, non-individualistic analyses are a basic requirement for arriving at intentional descriptions at all. In the case of rich representation, these teleological descriptions will not only be supplied by evolution or organismic features (although these do form a basis for describing the cognisers whose minds are endowed with rich intentionality), but by further normative principles. In this section I will focus on how the neuroscientific investigation of such normatively shaped intentional states can proceed.

II.8.2. Getting a Grip on Normatively Shaped Cognition

Human minds are typically able to acquire skills which I have been associating with rich forms of representations, namely the ability to be moved by semantic properties. Among other cognitive skills, some characteristic for the human mind are learning languages, acquiring and contributing to explanatory theories, and being able to grasp the kinds of theories which underlie all kinds of representation, be they linguistic, aesthetic, scientific, or psychological. (And for all we know, many of these cognitive abilities can be attributed to non-humans as well, at least in proto- or attenuated form; compare I.7.5.) For this reason, the task of explaining the mind cannot be achieved without acknowledging that it is shaped by the normative forces delineating such skills.

However, the fact that there are norms which have a causal influence on our minds is itself a descriptive fact, and not a normative one. Therefore, we can integrate norms into descriptive explanatory models. The fact that mental content is broad rather than narrow also means that individual analyses of cognitive capacities, such as investigations of the brain, do not themselves have to carry the burden of reducing norms (see section I.8). That is: If content were individualistic – if having intentional states would not at all depend on environmental factors –, then analyses of individual cognisers would also have to be able to settle matters of content. More specifically, if having a brain is the foundation for being a cogniser, then analyses of individual brains would have to carry this burden. I have argued that this is not the case. Rather, investigating neural properties explains why and how individuals fit into an environment which outfits their individual cognitive states with intentional attributes. Here, we can speak of an interpretation: environmental context supplies a function outfitting individual cognitive states with intentional interpretations (akin to Ned Block's "mapping theory", see Block 1991 and section I.8.5).

In order for any of us to be able to learn the skills associated with intentional properties, the characteristics of the norms guiding these skills have to themselves be learnable. To be sure, this does not imply that we could internalise every potential application of a rule (which would require infinite cognitive capacities). Rather, it means being able to extrapolate (potentially infinite applications of) a rule from finite observations of such applications. Cognitively, it means creating appropriately associated input-output-pairs, i.e. identifying a finite set of characteristics of an input which warrant yielding a specific (kind of) output. Since we can only ever extrapolate the characteristics of both in- and output from a finite amount of observations, we are never safe from being led to believe that we have learned a given rule whereas future application shows that we have in fact wrongly extrapolated the norm governing its application (compare Kripke 1982: 8 ff.).

So, whatever the relevant norms themselves are, they cannot be facts intrinsic to the cognitive make-up of someone who follows them (and they cannot be determined by the latter). While it is true that, under idealised conditions, the functions implemented in our neural mechanisms could also yield potentially infinite applications, we could not possibly determine whether the norm which governs a cognitive function conflates with the external norm which has caused the former's acquisition. That is, even if we could find out which formal rule is followed in either instance, we would not know whether it conflates with the rule which is supposedly followed.[100] Since the cognitive function has been acquired by way

[100] Both Wittgenstein (PI §193 f.) and Davidson (1980: 255-259) invoke the tempting image of a machine which could be broken down, much like an idealised neuroscientific methodology could make transparent neural

of extrapolating a rule from a finite amount of instances, we can only ever check whether future applications contradict the previously extrapolated rule, but we cannot find out the one rule governing all future applications. (The same is true for functions which are not acquired through learning, but through evolution, and which are in this sense innate. This is because even those functions which are innate are acquired by evolutionary means based on a finite determination of characteristic "fit" between a species and its environment. For example, the association between the magnetic north pole and anoxic waters has been acquired as the basis for the magnetotactic bacteria's mechanism because it proved stable over a finite amount of time. That the species' characteristics proved fitting for this finite amount of time, which caused organisms to pass on these characteristics to its offspring, does of course not imply that it is evolutionarily fit for all future instances of environmental conditions, so we cannot say that *the* single defining norm governing the evolutionary fit has been implemented in the respective species' genes.)

Since intentional interpretations are not merely accidentally imposed on behaviour or brain states, but themselves constitute some of the causal determinants of characteristic behaviour or brain states, it is crucial for explaining the structure of the mind. For example, we only fear stock-market crashes because we have socially been taught to do so, and we are only interpretable as fearing stock-market crashes for the very same reason. (I will go into further detail regarding the shaping of cognitive architecture by way of norms in the following section.) In this sense, intentional ascriptions capture phenomena pertaining to agential behaviour more accurately than non-intentional descriptions.

From assuming this form of intentional realism and the fact that all actual cognitive rule- and norm-related processing or behaviour can only ever be based on extrapolation from finite application it follows that intentionally explaining such processing or behaviour is also a matter of specifying finite states or processes. That is, for certain cognitive functions to count as intentional it is sufficient for them to be traceable back to the external norm (which governs the reference-relation) as its proper cause, not for the external norm to somehow be intrinsic to the cognitive make-up itself. Therefore, the conditions under which brain states imply intentional states are specifiable.

Any attempt at fully naturalising intentional properties of neural states still faces the final verdict on what the norms guiding these themselves are. As I have argued, they are not intrinsic facts about cognisers; instead, they are external phenomena which causally shape cognition. Anyone keen on naturalising intentionality tout court should like to treat such

mechanisms for rule-following, thus unveiling the secrets about which rule is actually followed – and they both reject this notion as misguided.

norms as properties of the natural world, that is, as properties restatable in non-normative terms. While this aspect exceeds my present reach, I will briefly offer a potential scenario in which one such naturalisation comes to pass. In it, norms turn out to be a highly complicated web woven by communities of cognisers and their culture. Here, naturalisation comes to terms with the development and spreading of such norms through communication, and succeeds in specifying how individual cognisers, who are both recipients and relays of such norms, react to those environmental cues which make up the communication that constitutes the building blocks of such norms.

While it is conceptually impossible to restate norms as non-norms, we can treat non-normative properties as constitutive for norms, insofar as some non-normative properties are sufficient for guiding even those forms of behaviour which we describe as normatively guided. That is, the set of all environmental cues which can be described as shaping a mind so as to internalise a norm is just that: a set of environmental cues; and environmental cues can also be described non-normatively. I have pointed out that the functional norms governing our cognition are arrived at through extrapolation from finite occurences; and while we think of these occurences, as well as of some of our cognition, as being guided by such norms (whether these are universal laws of nature or functional aims), we may choose to view these norms not as inherent facts of nature and cognition, but of our descriptions of these; and that, under radically different descriptions, they eventually evaporate.

II.8.3. A Schema for the Neuroscientific Investigation of Intentional States

As has been noted in section II.3, behaviour can be explained both by the means of intentional psychology and the means of neuroscience: that is, both intentional as well as neural states can be invoked as causes of behaviour. As has also been pointed out, the two kinds of explanation are not interchangeable, since judgments of appropriateness or error are only possible under the intentional explanation. No matter which norm exactly governs the intentional state – whether it is a social norm, a norm of rationality, one judging evolutionary adaptiveness, or otherwise –, it is clear that no such norm governs an individualistically described neural explanation (see I.8.5).

It is sometimes held or implied that neural (or physical, or "natural") explanations are not governed by norms, or not dependent on norms, or not assuming or implying norms, *because* they are causal explanations (see footnote 51). This rationale is misleading, since

intentional explanation is also a kind of causal explanation: the attribution of a mental state causally explains the behaviour in question (see I.6.3). However, what is true is that the intentional explanation is characterised *not just* by causal relations, but also by normative relations. So, what we should say is that the neural explanation is not governed by norms because it is *merely* causal, whereas the intentional explanation is *more* than just causal (compare I.4.3).

To briefly recapitulate (see section I.6.2 for the details): the respective causal relationships are certain lawlike relations of the form FA → GA ("if F happens to A, then G happens to A"). Any such law quantifies over objects which allow for these generalisations, and these objects are typically called "kinds" (in the natural sciences usually called "natural kinds") – in this case, the set of objects designated by A, F and G. The law specifies a certain property or starting conditions to which A is subjected (namely F), under which G occurs to A. (For illustrative purposes, think of A being water, F standing for "being heated to 100°C" and G standing for "boiling".) Laws are not simply generalisations, but rather specifications of what properties of which objects can be lawfully generalised (or, using Goodman's terminology, which are "projectible"; see his 1983: ch. 3). They are typically taken to explain an event event by (1) treating it as an instantiation of a specific law, or "subsuming" the event under a law and (2) by incorporating the laws themselves into scientific theories (which may themselves be sets of higher-order laws; that is, the fact that water boils at 100°C is not merely explained by stating that "all water boils when heated to 100°C" but also by incorporating this special law into more fundamental laws, e.g. concerning molecular movement).

The relationship between behaviour and mental states is "criterial" (Dennett 2007: 74), insofar as it is subject to certain norms characterised by inferential relations which are employed in agential explanations (cf. Levine 1987: 250). For example, exhibiting angriness-behaviour may typically be caused by a frustrating event, but it is crucially also explained by stating or supposing that it is appropriate to show this kind of behaviour in the case of such a frustrating event. So, it does not merely follow causally from a sparse description of the frustrating context (i.e. of its description in non-intentional/non-psychological terms) and its effects on agents, but also because there is an appropriateness-relation between such kinds of contexts and the respective kind of behaviour (see I.6.4). Such a relation singles out certain contexts as justifiedly or rationally eliciting angriness-behaviour, specifying the requirements for interpreting certain behaviour as intentional (see I.7.4). The respective contexts are commonly marked by psychologically loaded terms pointing to appropriate reactions, such as

"frustrating" or "annoying". To give another example: Sam's believing that Hugh is unmarried when being told that Hugh is a bachelor is explained by Sam's knowing English and knowing what a bachelor is. In both examples, the norms underlying or explaining the behaviour in question can be descriptively construed as causes for psychological dispositions or cognitive structures. And while the capacities employed in either example are at any given time physically realised, i.e. neurally implemented, they are only implemented the way they are *because* there is a functional relation between the frustrating context Max finds himself in and his angry reaction, and *because* Sam has, at some point, learned English and learned what being a bachelor means (see II.3).

Accordingly, any investigation of intentional states in neuroscience needs to account for the criteria delineated by the practice of intentional attribution. We know that raising one's hand does not by itself imply having any (or any specific) intentional state (cf. Danto 1973: ix f.), and therefore, neither can a brain state which can merely be described as, say, causing someone to raise their hand (compare Block 1995b: 398). Rather, the intentional ascription depends on the context of raising your hand: Its external cause and its relation to other intentional states (compare Dennett 1987: 93). Just as any intentional implication of the neural state which causes hand-raising depends on its external cause or its relation to other states which fall under intentional interpretations.

So, while neural descriptions can explain behaviour for which there is also an intentional description, they can only do so against the backdrop of said functional relations (such as a practice in which people learn appropriate behaviour, or languages, or semantic relations). If there were no such functional relations, there would also be no cognitive mechanisms for neuroscience to explain. Thus, it is generally false that neural and mental descriptions are merely different kinds of descriptions of the same thing: Neural descriptions are individualistic, mental ones are not. Therefore, mental descriptions take facts into account which strictly neural ones *cannot* take into account by stating (or implying/requiring) that what is described also falls under a norm. Since the two descriptions thusly refer to different facts, they cannot have the same object. (Again, this is not to deny that there is intentional behaviour which need not depend on an individual's learning to conform to a norm; but even insofar as it can be described *as* intentional, it is governed by norms, and thus construable as a relation between a cognitive agent and whatever her cognitive makeup is or has been shaped by. No such relation is implied by a neurobiological description.)

These inferentially characterised psychological properties (cf. Cummins 1983: chapter 2) can also be described as functions yielding correct outputs from given inputs (see II.5.2,

also compare Gladziejewski 2015). The in- and outputs are interpretable semantically, because only then can they qualify as meeting the applicable norm, as being correct or incorrect (cf. Searle 1983: 10). As mentioned, there can be different kinds of norms when dealing with matters of intentionality: evolutionary (innate) ones, socially acquired ones, and those concerned with basic matters of rationality. The following schema for mental functions applies to *all* intentional mental states, no matter their origin (socially imposed/learned or innate):

Schema M: [INPUT-signifier)] –FUNCTION→ [OUTPUT-signifier]

Every M can be stated by exclusively invoking physical (or chemical, or neural) terms. All that is required for them as pertaining to mental states, or to see them as normatively governed, is that inputs and outputs are *also* semantically interpretable. But this just relies on the classic idea of what a signifier is: a material object which has a semantic interpretation in a given symbolic system. It should be understood that schema M allows for a wider range of signifiers than the Saussurean notion (see section I.4.2): here they can be more than just symbolic signifiers (such as utterances, graphems and the like); rather, they can amount to any in- or output which warrants an intentional description (such as salient stimuli, objects of perception, and actions). By way of the signifiers' semantic properties – i.e. the fact that the input- and output-states are subject to semantic interpretation –, the functions underlying such schemas are connected to intentional mental states. In other words, by placing a causal physical process in an accurate normative context which treats the initial and resulting physical states as signifiers, these physical states are mapped onto intentional states and vice versa. This way, physical systems can be interpreted as "intentional machines" or "semantic engines" performing computations over semantic content (see I.4.4). However, nowhere *inbetween* in- and output (i.e. "internally" or "intrinsically") do we require anything to be interpretably semantically (although we can allow for it to be, as in the case of modules or compilers, which we could interpret as producing an intermediate semantic output and/or being sensitive to semantic input; cf. Levine 1987: 260).[101] Any physical process which

[101] Disregarding this fact has been a dominant root of confusion in (theoretically or actually) trying to map mental onto neural states, only to be bewildered by losing intentional properties in the process. By accepting that the states which are literally "inside our heads", and which enable us to have contentful states, not need have content themselves, we bridge the gap between semantic states and non-semantic states. Jacobson calls the latter "Aristotelian representations (…) [which] do not have content or satisfaction conditions" (Jacobson 2013: 45), suggesting that this is the kind of representations dominantly invoked in neuroscience.

causally derives one signifier from another is potentially subject to physical (or, in our case, neurobiological) investigation and in this sense "naturalised".

Signifiers need not stand on both sides of the schema. We know that some mental properties are effects of non-mental properties and vice versa (which, combined with the assumption of the causal closedness of the physical realm, has been a main motivator for wondering whether mental properties are in fact epiphenomenal, cf. Kim 1993: 280 f.). For example, lightrays emitted from an object X impinging on an agent's retina can cause her belief that there is an X. On the other hand, someone's desire to drink can cause the movement of a glass. So, we can expect to encounter processes which fit into either of the following sub-schemas:

Sub-schema M_{S1}: [INPUT-non-signifier] –FUNCTION→ [OUTPUT-signifier]
Sub-schema M_{S2}: [INPUT-signifier] –FUNCTION→ [OUTPUT-non-signifier]

For illustrative purposes, consider a student being tasked with giving a presentation on the mind-body-problem by her professor. If she succeeds, the student will eventually come up with behaviour which qualifies as giving this presentation, and we can not only assume that she eventually does so because she has learned to interpret her professor's utterances as a request to do so, but also that there is a causal process going on in her brain which leads from an initial physical state caused by the request to the eventual behavioural output when giving the presentation (see Figure 7 for this example and Figure 8 for the general schema). Given the information about the professor's request and some basic assumptions about the student's psychological make-up, there is an intentional explanation for her eventually giving this presentation, and the fact that she does so provides solid evidence for her having some specific intentional states (such as her prowess at the language she converses in with her professor, her knowledge about what giving a presentation requires, her motivation to do so, and so on).

Given this schema, it is easy to see how norms can shape cognition, which in turn produces behaviour that justifies intentional ascriptions. It is in learning processes that "a representation of the rules (…) [agents] follow constitutes one of the causal determinants of their behavior" (Fodor 1975: 74, fn. 15; compare I.4.4 and II.3), insofar as a physical process is selected (or "conditioned") which serves as an implementation of the function. Simply imagine you have several physical processes $F_{1\text{-}3}$ which yield different outputs $OS_{1\text{-}3}$ based on an input IS_1:

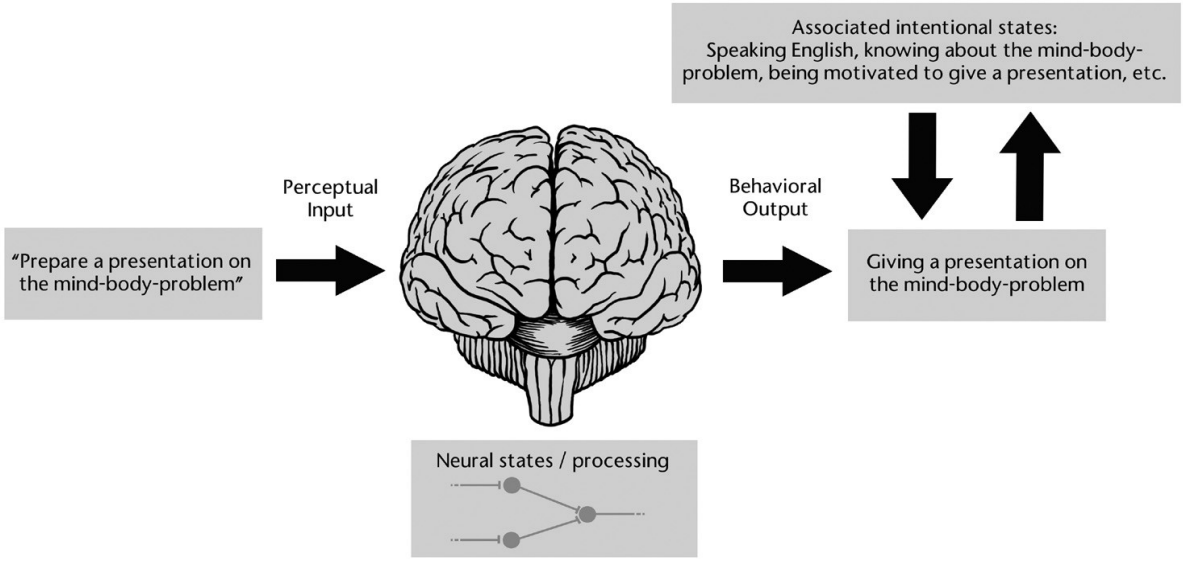

Figure 8: Schema M (example).

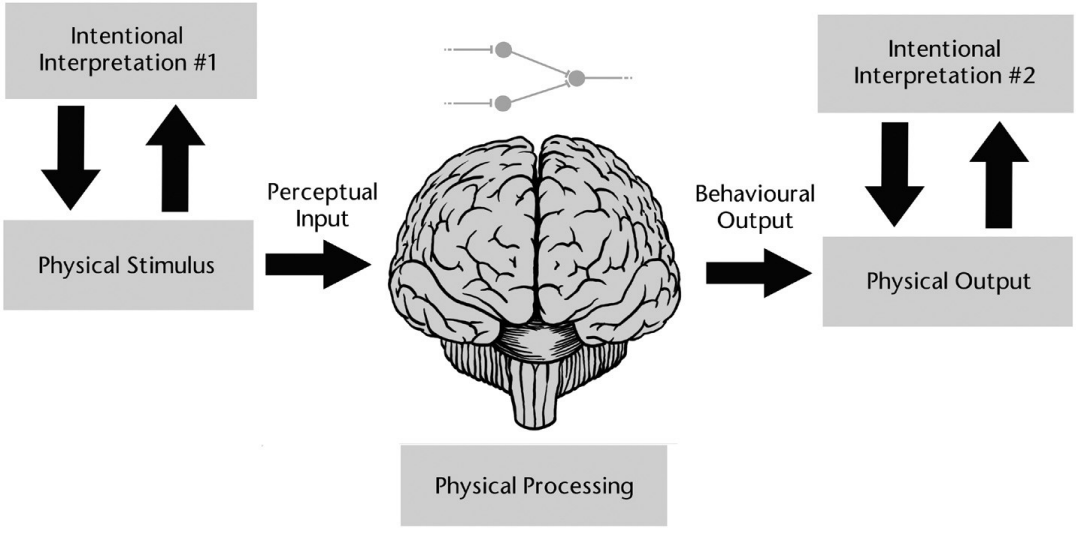

Figure 9: Schema M (general form).

$[IS_1]$ —F_1→ $[OS_1]$
$[IS_1]$ —F_2→ $[OS_2]$
$[IS_1]$ —F_3→ $[OS_3]$

Let's assume that out of these processes it is only F_1 which yields the output that corresponds to what is specified as correct or justified by intentional law. So, it is only when an organism computes F_1 that it is interpretable as correctly or justifiably having an intentional property pertaining to the respective law. And the way a "representation of the rules" causally influences behaviour is by selecting the suitable function F_1. (It should be easy to see how this general picture can apply to many different learning scenarios which range from learning in childhood to ongoing learning processes in adulthood, from learning languages and all kinds of social conventions to learning to operate tools, to the acquisiton of knowledge and so forth.)

Yet, representations are not themselves causally efficacious in the function's actual computation (i.e. in the process which leads from IS_1 to OS_1), since these are only the material parts of the desired symbols, and to assume otherwise would amount to committing the homunculus fallacy: to assume that the neural function is computed by a "homunculus pulling a volume off the shelf" (Levine 1987.: 254, compare II.3). Certainly there is no internal "list" or "rulebook" which is literally consulted by a brain-part: no neural process alone implies an intentional description, since, as just pointed out, the intentional description takes more facts into account than strictly neural ones. Crucially, the physicochemical description of a neural process is already sufficient for explaining those properties of an agent whose subset is potentially intentionally interpretable (i.e. it explains the totality of an organism's activity, and some of this activity can under certain internal and external conditions be interpreted as meaningful). A list or rulebook can be a causally efficacious factor insofar as it is part of the learning process, weeding out possible neurally implemented functions. This is the behavioural "trimming" implemented by the neural "pruning" we spoke of in I.8.5. Much the same goes for sparse forms of intentionality, in which it isn't representational properties which cause this trimming, but evolution – only that in such cases, neural functions are not weeded out within an organism's lifetime, and also not necessarily for social reasons.

To conclude this subsection I am adding a cursory list of what we can expect the description of the physical process whose in- and outputs are potentially intentionally interpretable to be like. (This is only a very humble hint based on current methodology and the research paradigms on which such descriptions depend, and it would have to be expanded

on by further experimental paradigms investigating correlations between brain processes and intentional states.) This list specifies and breaks down what I called an "intrinsic" description of an agent's properties in section I.8.5. Note that the numbered steps on this list do not necessarily indicate temporal sequences or ontological differences, but different steps in an analysis. For example, any signal transduction in the brain (or, at the very least, part of its sequence) will be identical to a change in the transducing neurons' properties, but we can still describe signal transduction and neural changes independently (and can analyse one in terms of the other). If we knew all the laws governing neural changes depending on signal transduction, we may be able to level this difference by reducing one to the other.

Note also that what I am labelling "signal state #2" is a pattern of nervous activation caused by "signal state #1", which in turn is an electrical and/or chemical effect of external causes by way of perception. This signal is a sequence of dynamic electrical patterns of which each subsequent state is an effect caused by the prior state and the properties of the physical objects transducing the signal (chiefly among them cellular and neural transmitter properties), and it causes neural changes (by strengthening used connections and weakening unused ones, cf. Hebb 1949) and potentially further bodily changes (by, say, innervating muscles).

1. Perception [signal state #1]: the effects of external causes on an organism's receptors
2. Electrical and chemical activity [the dynamic part of signal state #2]
3. Spatial and physical neural structures [the stationary aspects of signal state #2]
4. Changes in conductive properties [such as by Hebbian learning]
5. Behaviour [Motor responses caused by the transduction of signal state #2]

II.8.4. Translating Mental State and Neural State Descriptions

II.8.4.1. Requirements for a Translation

In the previous section I concluded that it is generally false that mental and neural state descriptions refer to the same things, since neural ones are individualistic whereas mental ones are not. Due to requiring that intentional states need to conform to a relevant norm, being in an intentional state can allow for many different implementations of this state. So, human beings with radically different brain structures could share the same desire, robots could share

specific beliefs with humans, and so on. Accepting this so-called "multiple realisability" view follows from adopting a kind of *functionalism* regarding mental states (see II.5, Fodor 1974, Greene 2015, Nathan & Del Pinal 2015: 5.2). For example, we thusly accept that if it is true that Max performed the addition task "2 + 3 = 5" yesterday at noon while his brain was in state S, then it does not follow that if Max's neural state at noon had not been S he would not have performed this task. Many other states which are not identical with S are potentially consistent with Max's performing an addition task. And the neurobiological description of S is itself not governed by the norm which specifies that what Max does counts as performing addition.

However, even though it is not *generally* true that mental and neural state descriptions are coreferential, they still can be so *specifically*. That is, it might be true of beings that have brains, or specifically of human beings, that only those who are in neural state S in fact perform addition tasks (or another intentionally characterised function). On the one hand, this may turn out to be a fact about our neural make-up: It could turn out to be true that all neural states which are not S and which humans are capable of having (or which they actually have) do not allow humans to perform addition. That is, any other neural state which would be interpretable as allowing someone to perform addition turns out to not be had by humans; or, in other words, out of the set of neural states humans are capable of having it is only S that enables them to perform addition. (Here we can distinguish between neural mechanisms in one person as opposed to one shared by several, so that translation could be feasible for neural states of one individual, but not for several.) On the other, it may be a truth about our environment: The norm which allows for the behaviour caused by neural state S to be interpreted as evidence for intentional mental state M might exclusively hold whenever S is realised, so that all neural states which coincide with M are S. Such conditions constitute the requirements for a *translation* between intentional mental and neural states. And whenever I talk about translation being feasible under some conditions, I mean conditions under which these requirements are met.

Whether they are met, and for how many intentional or neural states they are met, is under ongoing investigation by cognitive neuroscience, and we should interpret studies which relate mental and neural states as partaking in this endeavour. The kinds of facts we should expect neuroscience to deliver, which would impact such a translation, pertain to what kinds of neural states humans (can) have that underlie intentional capacities, as well as the brain's capacity for multiply implementing such states (for example, by cortical reorganization of shifting functionality from one region/structure to another, cf. Shih & Cohen 2004). Consider

once again the mindreading experiments discussed in section II.4, which seek to relate neural states and meaningful states. While the results of such studies are "only" of a probabilistic nature (but see footnote 78), these probabilistic results can still translate to the intentional description "it is more likely by such-and-such a degree that the subject has intentional state A than intentional state B (or C, or D, etc.)".

Settling the question whether intentional states can be inferred from neural states at all is a matter of figuring out the respective algorithm(s). We know that such algorithms do exist independently of our methodological access: For example, the correlation which mindreading decoders exploit in order to infer semantic content from neural states had already existed before the decoders were invented, and they would even have existed if such decoders had never been invented. As argued in section II.4, assuming that the agent herself has access to the information which these decoders extract, it is a truism to suppose that an algorithm which carries out the function in question is in fact implemented in the agent's brain. So we can assume that at least one algorithm which carries out the respective function is computable by a biological neural system (cf. Marr & Poggio 1977). However, as far as questions of methodological attainability go – whether we can decode the algorithms in question –, the jury's still out. This is one of the senses in which approaches that already assume reducibility of intentionality on neural states rely on "presumptive theses way out in front of the empirical support they require" (Dennett 1991b: 51).

Whether any such neural mechanism can be identified as producing outputs interpretable as evidence for intentional states is also tied to its being generalisable. In their mindreading study, Kay et al. point out that "[t]o be practical our identification algorithm must perform well even when brain activity is measured long after estimation of the receptive-field models" (Kay et al. 2008: 354), and that their results "demonstrate that the stimulus-related information that can be decoded from voxel activity remains largely stable over time" (ibid.).[102] While, given what has been said so far, we can at any time assume a token-token-identity between neural activity and the cognitive mechanism(s) underlying any intentional state, what makes intentional states inferrable from neural states is a certain temporal and structural stability of neural states correlatable to intentional states, or, in other words, we should require a relation between some form of mental type and neural type in

[102] "To assess performance over time we attempted identification for a set of 120 novel natural images that were seen approximately two months after the initial experiment. In this case 82% (99/120) of the images were identified correctly (chance performance 0.8%; subject S1, repeated trial). We also evaluated identification performance for a set of 12 novel natural images that were seen more than a year after the initial experiment. In this case 100% (12/12) of the images were identified correctly (chance performance 8%; subject S1, repeated trial)" (Kay et al. 2008: 354).

order to establish a mapping (cf. Nathan & Del Pinal forthcoming); with the most straightforward kind of neural type being a structurally or physiognomically defined one, namely in terms of arrays of kinds of neurons (and the properties related to their activity). Since any neuron instantiates a physical kind, it is described in terms of projectible properties, and thus in the form of a type: What any neural token can do in virtue of its being a physical kind, any corresponding type can do. The connections and structural relations between neurons ("networks") are also defined this way, and they can serve to develop a "syntax" of neuronal types: different combinations of different kinds of connections can constitute a combinatorial syntax insofar as we can build complex structures of neurons out of simpler "building blocks" and infer the properties of such complex structures from the properties of their building blocks (combined with appropriate connection rules; compare my discussion of compositionality in section I.4.4).

Again, whether we can come up with a suited compositional picture of the mechanisms underlying cognition is ultimately a question of disovering the nature of actual neural ensembles in the brain and their functional and structural stability over time. It is also a methodological question insofar as sufficient knowledge of neural organisation and development, combined with a suitable technological method, might allow us to track characteristic changes in neural organisation with sufficient precision over a relevant amount of time, so that a later neural state which structurally differs from an earlier one can still imply an intentional state just as reliably as the earlier one. Both issues of methodology and neural characteristics are open to discovery.

Also, characteristic impairments related to lesions are sometimes stated in intentional terms. For example, it has been claimed that damage in the ventromedial prefrontal cortex (VMPFC) leads to systematic deficits in moral motivation (cf. Roskies 2003). Here, the VMPFC damage itself is a neurological description, whereas the ascription of moral motivation is based on evidence related to the respective intentional state, such as self-reports (usually obtained through questionnaires), anecdotal evidence (such as biographical reports)[103], standardised behavioural evidence (such as the Iowa Gambling Task) and indirect physiological evidence for lack of motivation (namely, measuring subjects' skin conductance response). While results of lesion studies usually tell us which regions are necessary for having intentional capacities, but not which ones are specific to them, these can serve translations in a heuristic or cumulative way and add to other more specific results.

[103] The most popularly invoked example is Phineas Gage, a US railway worker who, following an accident in 1848, is taken have suffered from this kind of neural and behavioural impairment (cf. Damasio 1994: ch. 3; for criticism of Damasio's interpretation see Kihlstrom 2010 and Schleim 2010: chapter 3.2).

What is not under investigation by neuroscience, however, is whether the norms which specify that certain behaviour is justifiedly used as evidence supporting ascriptions of the respective intentional state which the neural description is meant to translate to do in fact obtain. Such questions fall to the sciences adjacent to the neuroscientific experimental methodology, such as psychology, philosophy, linguistics, anthropology, and so on. Experiments in cognitive neuroscience depend on there being an operationalisable psychological construct to investigate, and this operationalisation cannot itself recur to neurobiological constructs. For instance, our assumption that participants in the mindreading experiments perceive movies showcasing aeroplanes is based on what we take these movies to represent and that participants are versed in this kind of interpretation (i.e. we treat them as knowing that what they see are aeroplanes). While mathematical descriptions of the imagery they are presented plays an important role for coming up with a translational algorithm, it is readily acknowledged that its semantic aspects are also invoked in certain processing stages and contribute to the algorithm (see footnote 76). If we aim to completely subtract semantic aspects in our operationalised constructs (and disregard related semantic encoding in the brain), then mindreading experiments could be "narrow", insofar as they would only tell us about the processing of perceived contrast, colour, and the like. In this case, it may appear as if they could not be of any help to a translation from neural to semantic descriptions, but that is not entirely true. The fact that participants' perception is processed the way it is can be characteristically influenced by what participants expect or think they perceive (compare Rauss et al. 2011). Semantic knowledge can thusly play a role even for a subject's narrow properties; not necessarily in every brain region involved in such processing (again, compare footnote 76), and not in every act of perception, but certainly in some, and perhaps in those most characteristic for perception related to intentional ascriptions. Thus, we should expect even processing of asemantically operationalised constructs to be specific to the semantic content of the perception (and associated mental states or behaviour) and to aid in building a translation function.

Delineating the conditions under which a change in representational norms occurs which can impact the validity of a translation is much like asking "do aeroplane-symbols still refer to aeroplanes"? This kind of question is indigenous to matters of translation. Translation manuals change depending on whether certain words are still used to have the meaning they had when the manual was created. Consequently, the investigation of such matters will be integral to operationalising experimental constructs which are supposed to capture semantic

content (and once we have established translational bedrock, idiosyncratic understandings of certain concepts may even explain idiosyncratic processing exhibited by some subjects).

II.8.4.2. The Methodology of Translation

A translation is a way of establishing semantic relations by non-semantic means, and by not presupposing any semantic knowledge at all. Establishing a translation between mental and neural state descriptions circumvents semantic reasoning about the mind-brain relationship. The most popular (if not all) "philosophical" views about the mind-brain-relationship are based on semantic claims, some of my own included. For example, my conclusion that mental and neural descriptions cannot refer to the same things is based on semantic considerations (namely that mental content is broad), much as, say, Bennett and Hacker's argument that mental states are ascribed to persons and not to brains is (cf. Bennett & Hacker 2003: chapter 3). The latter can be trumped by adopting Dennett's brand of pragmatism (cf. Dennett 2007: 87) which holds that ascribing mental states is justified if such ascriptions prove useful, and why should we care too much about what people generally do when applying such ascriptions anyway? These are all examples of arriving at truths about the relation between mental and neural states by analysing the use of concepts. The downside to invoking such arguments is that opponents may simply question the argumentator's grasp of meaning. But until we come up with a non-semantic way of establishing semantic relations, such arguments will be the grounds on which we characterise the mind-brain-relationship.

It may seem a bit paradoxical that we cannot invoke translation as a non-semantic way of establishing semantic relations without accepting some semantic claims in the first place. If there was a way to start with a translation, then my arguments regarding broad content, much as most of my characterisation of mental state ascriptions in the first chapter, would be moot. However, it should become clear any moment now that my characterisation of mental states as theoretical explanatory terms, as crucially observable and intersubjective, as functional and their content as not essentially private, is what makes a translation which relies on correlating observable circumstances possible in the first place. So, in accepting the requirements for a theory of translation we do accept some semantic claims, but these are claims which imply that under some conditions semantic relations can be established empirically.

My proposed method of establishing a translation manual between mental and neural states proceeds along the lines laid out by Quine's "radical translation" and Davidson's spin

on Quine's ideas (also compare Lewis 1983b: 108-121). Essentially, radical translation is "[t]he recovery of a man's current language from his currently observed responses (…) [by] a linguist who, unaided by an interpreter, is out to penetrate and translate a language hitherto unknown. All the objective data he has to go on are the forces that he sees impinging on the native's surfaces and the observable behavior, vocal and otherwise, of the native" (cf. Quine 1960: 28). Radical translation is an empirical form of establishing semantic relations without being able to invoke any semantic knowledge.[104] Here, interpreting foreign utterances as meaningful depends on building a systematic, interdependent web of (what Quine calls "analytic") hypotheses about the dependency of a speaker's utterances on external circumstances (ibid.).[105] As described in I.7.4, the translational method consists in coupling (foreign) utterances which are assumed to express mostly true beliefs with true descriptions of external circumstances (expressed in a familiar tongue).

In Davidson's case, the idea that theories of meaning are based on gathering correlations is based on an intricate argument (in his 2001a: 17-36, see also 216 f. and 224 f. for summaries of his main points). Now, if we accept Davidson's acquisition of meaning by way of triangulation (see I.7.4), the picture itself should be clear enough: A learner systematically correlates a teacher's utterances with external circumstances in the world. So her web of hypotheses about the meaning of utterances will have the status of an empirical theory which is evidentially supported by observed correlations. Still, the way Davidson originally established gathering correlations as an empirical method for building theories of meaning is independent from his claims about triangulation (or rather, the latter emerged from the former). Therefore, I will briefly sketch Davidson's original argument. First off, any theory of meaning should meet two criteria: it should give the meaning of every possibly sentence of a language (due to semantic holism, see I.7.4), and it should do so without recurring to semantic notions. To achieve these aims, we proceed from semantic specifications of the form [1] to [2] (whereby we specify extra-linguistic conditions tracking the meaning of the quoted sentence which is to be translated) and finally to [3] (whereby we get rid of the intensional/semantic notion "means"):

[1] "Schnee ist weiß" in German means "Snow is white" in English

[2] "Schnee ist weiß" in German means that snow is white

[104] While they are not known and cannot be assumed, these semantic relations are based on objective correlations which *are there*, much as an ideal algorithm (i.e. the systematic relation which is the prerequisite for there being an actual algorithm) exists before it is actually discovered (see II.4).

[105] I should add that Quine's ultimate aim here was to propound his thesis about the indeterminacy of translation (see I.6.1). However, my focus is exclusively on his points about the methodology of a radical translation.

[3] "Schnee ist weiß" is true in German if and only if snow is white

Since they specify truth conditions, biconditionals (i.e. sentences using the connector "if and only if") such as [3] are called "T-sentences" (cf. Tarski 1986: 51-198). Theories which yield truth conditions for all possible sentences of a language are treated as specifying the *meaning* of the sentence on their left hand side. Essentially, Davidson defends this procedure in the following way:

> "How can a theory of absolute truth (...) be considered a theory of meaning? (…) The question to ask is whether someone who knows a theory of truth for a language L would have enough information to interpret what a speaker of L says. I think the right way to investigate this question is to ask in turn whether the empirical and formal constraints on a theory of truth sufficiently limit the range of acceptable theories. Suppose, for example, that *every* theory that satisfied the requirements gave the truth conditions of 'Socrates flies' as suggested above [i.e. in the form of a T-sentence]. Then clearly to know the theory (and to know *that* it is a theory that satisfies the constraints) is to know that the T-sentence *uniquely* gives the truth conditions of 'Socrates flies'. And this is to know enough about its role in the language.
> I don't for a moment imagine such uniqueness would emerge. But I do think that reasonable empirical constraints on the interpretation of T-sentences (the conditions under which we find them true), plus the formal constraints, will leave enough invariant as between theories to allow us to say that a theory of truth captures the essential role of each sentence. (…) I suggest that what is invariant as between different acceptable theories of truth is meaning. (...) Different theories of truth may assign different truth conditions to the same sentence (this is the semantic analogue of Quine's indeterminacy of translation), while the theories are (nearly enough) in agreement on the roles of the sentences in the language" (Davidson 2001a: 224 f.).

One initial problem, which this defense means to address, is that "Snow is white" is true if and only if 1 + 1 = 2, even though these two statements obviously do not mean the same (ibid.: 25 f. & 138).

> "[W]e might be misled by the remark that the (…) [T-sentences] could be read as giving meanings, for what this wrongly suggests is that testing a theory of truth calls for direct insight into what each sentence means. But in fact, all that is needed is the ability to recognize when the required biconditionals are true. This means that in principle it is no harder to test the empirical adequacy of a theory of truth than it is for a competent speaker of English to decide whether sentences like '"Snow is white" is true if and only if snow is white' are true. So

> semantics, or the theory of truth at least, seems on as firm a footing empirically as syntax" (ibid.: 61 f.).

This combination of holism (ibid.: 138 f.) and empirical testability of such truth-theories (ibid.: 135) ensures that if we know everything which is relevant for the truth conditions of every possible sentence of a language, then that is all we could or need to know about their meaning. For example, how can a competent speaker of English ever learn that "Snow is white" does not mean "1 + 1 = 2", if all she can ever do to find out about linguistic meaning is empirically gather correlations? Firstly, her hypothesis about what "snow is white" means is also holistically informed by her hypotheses about what "snow fell on Christmas Day" and "this dove is white" mean, and how the use of individual words influences the truth conditions of sentences they appear in. Secondly, she can observe external circumstances which provide plausible causes (or reasons) for speakers to utter either "snow is white" or "1 + 1 = 2", and these usually differ.

From Quine's and Davidson's theory I retain the form of specifying meaning empirically by way of gathering correlations (ibid.: 135). The parameters invoked in such specifications might differ depending on the kind of mental or neural state, and the best way of describing them; but for starters, I propose to characterise a specific mental state by specifying the intentional agent A, the intentional mode I, the propositional content P, a time index t and the relevant context for the attribution C, and to characterise neural states by specifying a nervous system N, an output O dependent on the applicable method [for example, in the case of fMRI, this would amount to a combination of anatomical descriptions and a superimposed/dependent pattern of activation], a method M, and a time index t. At a glance:

$[A_n, I_n, P_n, C_n]$ coincides at t_n with $[N_n, O_n, M_n]$

A systematic gathering of correlations should seek to eliminate time, person and context variables as far as possible. The degree to which these correlations can be generalised as to eliminate individual differences, or to factor out neural differences, will determine the extent of the resulting translation manual. Ideally, we arrive at:

$[I_n, P_n, C_n]$ *if and only if* $[O_n, M_n]$

Also, by varying intentional modes (attitudes) while keeping propositional content constant, and vice versa, we should attempt to separate these two factors and identify corresponding factors on the neural side. Something similar can be attempted by modulating operationalised construct and methodology independently; meta-theories about the nature of different methods of measurement, and the way results depend on each of them, will prove integral to this endeavor (compare section 3 in Sullivan forthcoming).

Since distinct cognitive tasks can involve the same neural structures, while the same cognitive task can be executed by distinct ones, substantially distinct mental properties can turn out to correlate with the same neuroscientific term and vice versa. In such cases the context variable will have to be retained to hold on to terminological differences in one language when we are unwilling to surrender them to the other's terminological poverty (despite creating additional queries, such as an adequate formalisation of context statements). That is, it can be obvious to us that raising one's hand in order to vote is strikingly different raising it in order to swat a fly, even though some of the underlying physical processes might turn out to be the same. If some intentional states turn out to be physically underdetermined (cf. Barrett 2006 & 2012, see II.8.4.5) – that is, if a set of distinct intentional states maps onto the same physical state – then we obviously need both the physical information *and* the context variable to distinguish between them. And, vice versa, if there is a set of possible physical implementations of an intentional state then we need additional information to infer the neural state from the intentional one.

If methodology permits, vocabulary of a translation manual may be enriched for practicality and/or explanatory value: If it is not practical to look for a neural description for each mental description, the latter may be broken down according to its syntax (such as intentional modes, types of content), or vice versa (anatomical details, concentration of neurotransmitters in certain areas, etc.). For example, it may turn out that we can track distinct mental states to distinct neural states, but not their individual terms. For example, the difference between the belief that snow is white and the belief that grass is green may be consistently trackable in neural terms, but maybe not the difference between having snow as the content of a belief and having grass as the content of a belief (and similarly for "is", "white" and "green"). Neural correlates for intentional terms or intentional correlates for neural "atoms" (such as types of neurons, transmitters etc.) may need to be introduced in order to limit ambiguity; but at the same time, introducing such terms depends on there being consistent correlations. So, there may not be an explanatory need for introducing them in the

first place; but if there is (due to the need for limiting ambiguity), then it may practically not happen for lack of consistent correlations.

Also, we could find "neural tautologies" which appear in correlations but are obsolete for translation. For example, one particular neural activation pattern N_0 may occur at all times. Additionally, assume that mental description M_1 is true if and only if neural description N_1 is true. In this case, the fact that N_0 and N_1 are true whenever M_1 is true does not mean that M_1 means N_0 and N_1. Rather, N_0 is a term which has no significance for M_1. Here, the activation pattern N_0 and N_1 needs to be split into two terms: One which means M_1 and one which does not. Something similar may apply vice versa, only that we must distinguish between beliefs which are true at all times (such as the belief that 1 + 1 = 2) and beliefs which are reasonably expressed under all circumstances (namely none).

Much as Haynes & Rees have stated (see II.4), what the reach of such translations track are the degree to which a distinction made in one theory neatly coincides with a distinction in the other. And since this distinction can track truth conditions, it is all we can and need to do to preserve meaning in a translation. However, the biconditionals we are using to track such distinctions may plausibly turn out to only support local translations which are considerably more restrictive than the term "translation" suggests. For example, the temporal and intersubjective stability expected from common translation manuals might turn out not to hold for the presently proposed mental-neural translations (also see II.8.4.4). They might not apply across individuals, across species or even across one single individual's lifetime:

> „A ["species-specific biconditional law"] states that any organism or system, belonging to a certain species, is such that it has the given mental property at a time if and only if it is in a certain specified physical state at that time. (…) In order to generate laws of this kind, biological species may turn out to be too wide; individual differences in the localization of psychological functions in the brain are well known. Moreover, given the phenomena of learning and maturation, injuries to the brain, and the like, the neural structure that subserves a psychological state or function may change for an individual over its lifetime. What is important then is that these laws are relative to physical-biological structure-types, although for simplicity I will continue to put the matter in terms of species. The substantive theoretical assumption here is the belief that for each psychological state there are physical-biological structure types, at a certain level of description or specification, that generate laws of this form. (...) Unlike species-independent laws, these laws cannot buy us a uniform or global reduction of psychology, a reduction of every psychological state to a uniform physical-biological base across all actual and possible organisms; however, these laws will buy us a

> series of species-specific or local reductions. If we had a law of this form for each psychological state-type for humans, (…) [it] would tell us how human psychology is physically implemented, how the causal connections between our psychological events and processes work at the physical-biological level, what biological subsystems subserve our cognitive capacities and functions, and so forth“ (Kim 1993: 273 f.).

Yet, at least as far as temporal stability goes, this fact sets our form of translation only gradually apart from linguistic translation: for instance, if I were to visit China, I would not bring a 200-years-old dictionary with me.

II.8.4.3. Lost in Translation

Contrary to what Kim says above, I reject the notion that such biconditionals support reduction, even while accepting all of his other stated claims. But why can't the resulting mappings be used as "bridge laws" (cf. Nagel 1961: ch. 11, Sklar 1967: 118-121) in a reductionist programme? Now, since intentional mental states are explanatory by way of requiring norms about the proper relations between cognitive mechanisms and the environment, subplanting them with descriptions which do not depend on the environment (or only in a considerably more limited way), *something* must get lost in translation. But how can this be, if translating by definition implies preserving meaning? It is because the correlations we can gather between intentional and neuronal states, which form the basis for a translation manual, are systematically bound to the environmental conditions under which they are gathered, namely by way of teleological principles which serve as methods for determining representational content (as described in II.7). Some of these conditions are constitutive for the meaning of the mental terms. If they change, the translation is moot.

For example, while establishing correlations, we might find some beliefs, each of which at all times correlate with a neuroscientific description belonging to the belief's holder; and these might be beliefs as straightforward as "I am happy right now" or "I am sad right now". Intuitively, most people, or at least those who aren't openly misanthropic, would hold that people should rather be happy than sad. However, that it is in our power to change the neuronal state correlating with the sadness-belief to the one correlating with the happiness-belief does not imply that we should, even accepting the norm that people should rather be happy than sad. Since there in fact *are* and have been many such cases, it is easy to see why:

Changing such states can yield dysfunctional agents, such as in the case of drug abuse. Leaving aside any considerations of side-effects of drugs for the moment, which can also provide reasons against using them, the reason is this: The norm that people should be rather happy than sad does not imply that people should be happy under all circumstances. This is because there are psychological laws which state the conditions under which being happy is justified/rational. Being happy over the death of a loved one is, ceteris paribus, not justified, and neither is being sad over a missed friend's eventual return. Only under very limited constraints can it be deemed rational to find beauty ugly, comfort abhorrent, or the like – namely, when someone does in fact accept enough related rational norms (compare Davidson 1980: 222). Special cases can violate laws of rationality, but they cannot be violated across the board. Rational norms need not be abided by for every singular ascription of an intentional state, but the fact that any single ascription can be made is based on there being laws of psychological rationality whose holding the ascription requires (see I.7.4 and I.8.5). Again, this does not mean that people cannot be inappropriately happy; it just delineates the requirements for us to ever perceive someone's happiness as inappropriate.

In the case of a purely neurobiological description, there is no such implication. Ideally – if we knew all the causes and consequences of neural activation – we could know the conditions under which a neural mechanism induces behavioural happiness-states. But whenever we want to find out about the appropriateness of a mental state, we will have to gather information that goes beyond just this neural state. And that is just to say that we will invoke the respective intentional law and relate it to the activity of the neural mechanism. In this sense, it is methodologically impossible to fully subplant intentional state ascriptions with neural state descriptions.

Yet, translations may appear to yield contrary results: If there are stable correlations between intentional and neuronal states, then they are translatable. If they are translatable, the meaning of the original and the translated term are identical. If their meanings are identical, we can use them interchangeably. Which seems to contradict what I just said.

The solution is that what is required for translation are actually stable correlations over (potentially) different environmental conditions. Once the environmental conditions change so as to violate the requirements of the respective law of psychological rationality, the translation has to be revised. But this is just a reminder that broad content depends on the environment while narrow content does not (see I.8), and that both can only be equated when relevant environmental conditions are held constant. A translation manual gets outdated when a sufficient amount of meaningful environmental properties changes, or, more specifically:

when the environmental conditions change so much as to completely obscure a given cognitive mechanism's functionality. This is not to say that semantic descriptions become moot whenever a mechanism becomes dysfunctional, but that, if enough dysfunctionality accumulates, it can obscure its semantic object to the point where we cannot even describe the mechanism as misrepresenting something. For example, when there are more toxic elongate objects moving parallelly to their longitudinal axis in a toad's environment than nourishing ones, the neural mechanism underlying its predatory behaviour has outlived its functionality (see II.6). It can still be described semantically as misrepresenting non-nourishing objects as nourishing ones, but only as long as we can reconstruct the mechanism's true purpose. As long as the intentional description has explanatory value, it is applicable. This is why a *sufficient* amount of environmental properties has to change: they have to change so much as to obscure interpretability. This is the narrow equivalent to what Davidson's criteria for interpretability in rich contexts are (see I.7.4): if irrationality pervades attempted explanations in intentional psychology, at some point the respective agent cannot even be conceived of as irrational any longer. Standards of rationality cease to matter when they cease to explain an agent's behaviour: the agent appears arational rather than irrational and stops being an agent.

II.8.4.4. Incongruencies between Mapped Kinds

Mindreading studies (see II.4 and II.8.4.1) have shown that a mental-neural translation is in principle feasible: Correlations between neural activity and perceptual content have proven stable enough to support inferences from fMRI-data to intentional states with a significant rate of success. While the translational algorithms used in these experiments are relative to individual subjects (and therefore do not yield general "translation manuals"), the method used to come up with these algorithms has proven to be intersubjectively applicable. So, while input-output connections differ across individuals, the functions for each are determinable by having each individual undergo a series of trials.

Algorithmic outputs in such studies are characteristically probabilistic (compare footnote 78). In the case of decoding moving images, visual outputs are superimposed according to probabilistic weights (cf. Nishimoto et al. 2011: figure 4). That is, even though subjects view novel images which the decoder hasn't been trained on, it is methodologically assumed that there is only a finite range of possibly viewed movies which can cause the neural activity: Based on the measured activity the "trained" algorithm determines how

probable each image out of the ones it's been trained on is and superimposes them accordingly. If this methodology were to work analogously for tracking propositionally individuated beliefs, desires and intentions, we would get results such as

if [agent A's neural activation] then [p_1 (A believes PB_1) and p_2 (A believes PB_2) and... p_n (A believes PB_n) and p_3 (A desires PD_1) and p_4 (A desires PD_2) and... p_m (A desires PD_m) and p_5 (A intends PI_1) and p_6 (A intends PI_2) and... p_o (A intends PI_o)],

where each p is a probabilistic weight, each PB is the propositional content of a belief, each PD that of a desire and each PI that of an intention.

A perceived weakness in the analogy between viewed images and held propositional attitudes might be that superimposed images form a visually graspable image (which shares visual features with the image initially viewed by the subject), while propositional attitudes cannot be similarly superimposed. However, superimpositions are just visual representations of conjunctions: so, the possible output statement that, say, there is a 70% chance that at a given time a given subject believes P_1 and a 20% chance that she believes P_2, then this statement plays much the same role as any visual superimposition in the cited study.

Once established, the probability with which such a translational algorithm yields a false result marks a lack of taking into account a decisive variable regarding the mapping. This variable may be a neural one (e.g. a neural activation pattern which is not fed into the algorithm – in the mindreading studies, only *some* brain areas are scanned, and perhaps "reading" additional areas could compensate for some of the resulting uncertainty; see footnote 76) or an external one (i.e. that the intentional property of a behavioural output depends on one or several factors which cannot be accounted for by "reading" the neural activation alone).

These and similar difficulties in establishing correlations between neural and intentional states are of varying relevance to translational mappings. Some are methodological in nature and should be distinguished from the claim that, even under ideal contextual and methodological conditions, mental kinds do not map congruently onto physiological or specifically neural ones (see Figure 9). Now, since mental categories depend at least implicitly on taking external factors into account (namely because they are individuated by their content, and content itself is determined not by internal or intrinsic factors alone, see I.8) whereas neural ones do not, neural and mental state descriptions *never* map neatly. Still, we can give meaning to this claim when worrying that (1) several or all type-terms from one

vocabulary do not correlate with or map onto type-terms from the other but only with an arbitrary set of tokens, so that the translated terms would be wildly disjunctive (cf. Fodor 1974: 103 f.), or that (2) no matter whether they map onto types or tokes, they never correlate systematically enough for there to be a mapping-algorithm outputting significant implications.

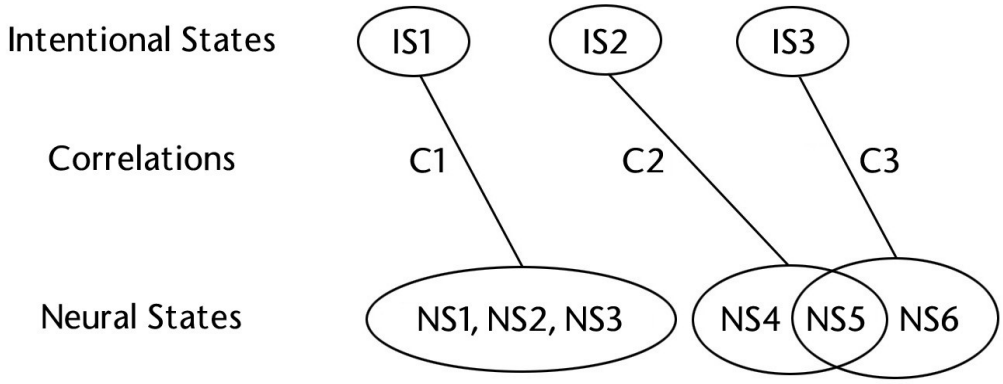

Figure 10: Although intentional states may not map congruently onto neural states, they may still be inferrable. If we interpret balloons as sets of conjunctions, then IS_1 is inferred from the conjunction of NS_1, NS_2 and NS_3, IS_2 from NS_4 and NS_5, IS_3 from NS_5 and NS_6, and vice versa. However, if we interpret balloons as disjunctions of states, then three different states NS_1, NS_2 and NS_3 are inferrable from any instantiation of IS_1. Such cases of ambiguity are unproblematic if each of the three neural states maps back to IS_1 and/or if they are resolvable by invoking contextual information: say, whenever the neural states are embedded or embeddable in a context which distinguishes between the associated intentional states. Problematic cases are those where ambiguities cannot be resolved and no use of a kind term from one theory distinguishes between several mutually exclusive kind terms from another theory (NS_5, which maps both to IS_2 and IS_3, might be such a term).

The second point can be dealt with comparatively briefly, so I will consider it first. Again, the worry is that, if different instantiations of the same mental state do not sufficiently coincide with the same neural state or vice versa, there can be no (significant) correlations. This worry has two major roots, namely the idea that from accepting functionalism about mental states it follows that they cannot be (or are unlikely to be) systematically tied to their "multiple realisations" or implementations (cf. Kim 1993: 273-275, 309 ff., 341 f.). Yet, we have to presume no such thing. A lack of systematic connections may turn out to be true empirically, but there is no conceptual reason to conclude the impossibility of translation from functionalism alone. As I have stated at the outset (see II.8.4.1), for translation to work it is required that the different physical realisations of mental states – manifold as they may be –

are de facto finite, systematic and can be specified. And even though functionalism says that we could dream up an infinite amount of physical realisations for any single mental state, it is not implied that the realisations that *do* exist are infinite.

The second root of this worry is that brains could be too diverse and malleable in individuals, in or between species in order to tie neural properties systematically to mental states (compare II.8.4.2). But, similar to our retort to its first root, recognizing phenomena such as neural plasticity, differing physiognomies and onto- and phylogenetic histories, convergent evolution and the like does not commit us to ruling out that mental states are still systematically and trackably realised. For example, although the wings of birds and bats do not share the same physiognomy and evolutionary history, they both serve the function of flying, and we can track how either physical implementation serves this function. This is because this and similar connections between function and physical implementation are generalisable *enough*, and so what we need to find out about mental states is whether there actually are enough systematic realisations of them as well. In fact, we have good reason to expect that there are such trackable systematicities between mental states and their realisations to be found, since, on the one hand, the evolutionary success of any organism depends on the systematicity of the connections between its brain, behaviour and environment and, on the other, mental states track determinants of behaviour in relation to environmental context. And while it is reasonable to assume that brains are malleable, organisms would be in poor shape indeed if their brains changed *arbitrarily* in a way unrelated to their mental functioning. Rather, neural structures need to change somewhat systematically, namely so as to be favourable in regard to behaviour and environment. Some mental states may turn out to be trackable across species, some across populations, some across individuals, and some merely across time within *one* individual (cf. ibid.: 274 f.). In each of these cases, we can establish mental-neural translations with differing reach, and if no such case is to be found, then translation does not come to pass – simple as that. (Another question, which I am not pursuing here, is whether current or future neuroscientific methodology is up to the task and whether scientific practice is actually aimed at establishing this kind of stable mapping; cf. section 3 in Sullivan forthcoming).

So, what about the remaining worry, that translations could be inadequate due to their potentially mapping types from one theory onto disjunctive terms from another? If through correlative experiments we were to find out that the instantiation of a mental kind-term *never* maps onto the "same" neural state (i.e. one which is reasonably similar by a neural criterion), we would end up with a mapping of a mental type to different neural tokens at different times

and no correlation at all. Saying that the mental state maps onto the disjunction of all these singular neural tokens is obviously a terrible move, since the disjunctive term is completely useless for any attempt at translation due to its not supporting counterfactual inferences to mental states. However, construed this way, this worry conflates with the worry I just dealt with: that there simply turn out to be no systematic correlations to be found. In order to distinguish the two worries, we need to construe this one a bit differently, namely as holding that a translation may be inadequate due to its mapping a mental state onto a disjunction of heterogeneous neural *types*, not tokens. (And assuming that this disjunction is not a neural type itself.) So, this is the worry I will be dealing with in the remainder of this section.

It is crucial to note that the mapping-algorithm's output states are not meant to actually *replace* kinds in the theory which provides the input, but only allow us to infer one from the other, given some contextual conditions which preserve the explanatory differences between our two theories (i.e. the conditions related to the semantic broadness of one and the narrowness of the other). Therefore, the principal point that disjunctive kinds may not be exploitable by bridge laws since they themselves do not constitute kinds in their indigenous theory is moot (cf. Kim 1993: 317 f.). Simply put, the required mapping function is weaker than a bridge law. Since it only needs to support inferences from one kind of state to another, given that contextual explanatory factors are held constant, a mapped term's being "wildly disjunctive" would by itself not be problematic for our endeavour (cf. Kim 1993: 316-319). Rather, we require that translational algorithms map to types in a non-ambiguous way (see Fig. 9). So, its mapping to disjunctions of types is only a problem when ambiguity ensues. That is, translation can work if A maps to B or C and both B or C *exclusively* map to A. Here, B and C could be multiple realisations of A, but for translation to work well either of them needs to *only* realise A (much like the wings of birds and bats both realise flying, but not swimming).

We could allow for *some* translated terms to be ambiguous (as is also usually the case between natural languages), but certainly not for all. For example, German contains some ambiguous words like "bank", which translates both to "bank" as well as to "bench" in English. Yet, it is still translatable, since we can invoke contextual information to clear up ambiguities. A translation can only be complete if for every ambiguous term there are enough terms related to it which are not ambiguous and which can be invoked for clarity. And even when there aren't, leaving a linguistic fragment untranslated need not be so catastrophic as to undermine a translation's explanatory power on the whole. But if enough of the invoked context were ambiguous too, we would be utterly lost. In our case, the overall threshold of

when translations break down, signifying the loss of all explanatory power, ultimately amounts to the value of statistical significance related to the hits and misses of the translational algorithms.

So, there is no need to insist that whatever neural property correlates with a mental property is itself a non-disjunctive kind-term in neuroscience or vice versa. What is needed is that there is a systematic correlation between *any* set of neural properties and a large enough set of intentional state terms yielding a computable algorithm which tracks differences between intentional states by tracking neural differences and vice versa. Each side of individual correlations might very well be wildly disjunctive; and here, "wildly" can only mean "unwieldily" (as in: unlike an intentional attitude ascription, the corresponding data derived from neural activation fed into the algorithm would take any human being a while to read out loud) but not "unsystematically". That is, they can be disjunctions in terms of neural/physical kinds, but they cannot be disjunctions of neural states each of which would correspond to a different mental state so that the entire set of implied mental states were irresolvably internally inconsistent. In such a case, translations would be too ambiguous to be explanatory. While some ambiguities can be tolerated and resolved on the whole, they need to be kept in check.

II.8.4.5. Kind-Revisions

In this final section I will exploit some of the lessons learned in the previous one and apply them to a current debate in theory of science. The question driving the debate is this: Could or should incongruencies between mapped kinds lead to revisions of kind-terms of either of the two theories which supply these kinds, specifically in cognitive science? For one, we do assume that there are close relations between what happens in the brain and what happens mentally, so the idea that theory-formation in psychology could or should generally have an effect on formation in neuroscience, and/or vice versa, presents itself. Consequently, the view has been prominently advocated that, if mental and neural theories turn out to not neatly map onto each other, nomological categories in one field should be revised in light of the explanatory properties of the other categories. And, since neural kinds are generally viewed as the ontologically more basic or "natural" ones, the dominant view is that it is the mental categories which should be revised (or even eliminated) in favour of neural categories. In the following, I am only going to discuss revisions, not elimination, since eliminative

theories generally assume that intentional theories have no actual explanatory power at all, or don't refer to anything "real", which is a problematic assumption I do not share. (For the most prominent view on elimination, see Churchland 1981 and 2005.) My disregarding eliminative views notwithstanding, what to do when relating mental and neural categories turns up both commonalities *and* discrepancies is in fact a pressing issue, since, beyond expecting an overall covariation between mental and neural goings-on, there is no theoretical reason for expecting translations to turn out so smoothly as to amount to a simple one-on-one mapping.

First, let me point out when revisions are uncalled for: namely, whenever mental and neural kinds constitute distinct explanatory factors within one theory. For example, there are ongoing attempts to trace psychiatric disorders back to neural causes. However, since "[t]oday, most clinical psychiatrists try to understand mental illness in the conceptual framework of a so-called "bio-psycho-social" model that integrates the action of biological, psychological, and social variables, and helps to design personalised therapeutic programs" (Tretter et al. 2010: 31), both neural and mental kinds are marked as having a causal effect on the respective theory's kinds (i.e. as variables contributing to the explanation). Here, "revising" can only mean optimising the integration of neural variables into the theory, not actually revising the non-neural kinds in terms of the neural ones. (I will leave open the question how such disorders themselves qualify as "mental kinds"; but, as I have pointed out in section I.9.1., we do retract agential ascriptions in such cases, so mental disorders or their symptoms cannot themselves amount to paradigmatic intentional states. For a broad discussion of whether psychiatric disorders amount to "natural kinds", see Kincaid & Sullivan 2014.)[106]

Generally, the idea is that so-called "natural kind revisions" proceed by aligning causes between different theoretical categories, insofar as these categories

> "ought to group together phenomena in such a way that they are subject to the same type of causal explanation (see, e.g., Craver 2009) and respond similarly to the same kind of causal interventions (see, e.g., Woodward 2003). If psychiatric categories do not find such groupings, there is reason to revise and/or eliminate existing classifications" (Kincaid & Sullivan 2014: 2).

[106] The mere fact that different kinds of "kinds" are contributing to an explanation does not imply that these kinds are in any way competing. Yet, this can be insinuated by sticking with the term "natural kind", rather than explicating kinds in terms of specific explanatory roles. If one kind is "natural", but another is, say, "interactive" and therefore "unnatural", then surely the "natural" one must enjoy primacy (see II.5.4)? But this is just to want to make a case against explanatory power having primacy in theory-formation, and there can be no such case. Sometimes (and perhaps in psychiatry, see Tretter's quote below), theoretical pluralism may just turn out to be the right way to go. Of course, we cannot know for sure until a supreme explanatory theory has been found to be workable; all we know for now is that in some cases, distinct kinds serve their explanatory role *by being distinct*.

However, given the standard notion that kinds and causal relations are defined in terms of each other, the two quoted sentences don't seem to fit together: kinds are terms whose meaning is (exclusively or at least primarily) defined by the laws they support, and, in theories which use the notion of causality for explanatory gain, laws are causal relations between kinds (see I.6.2). So, it would be inconsistent to require that kind-terms be revised in light of causal relations. For, firstly, either these causal relations are an accepted part of the theory which uses the kinds, then the kinds' meanings are already set by them and there is no need for revision. Or, secondly, these causal relations are rejected (i.e. the corresponding laws are found to not be generalisable), then they cease to be part of the theory and do not affect the kinds to begin with, which also does not invoke the need for revision. Or, lastly, the causal relations are part of another theory, but then they cannot line up if their kinds don't. So, the only way a theory's kinds can be "revised" is by its adopting true causal laws (although it would be more proper to say that in such cases it is the theory itself which is revised in light of its lining up with available evidence, not its kinds in terms of their lining up with causal relations). Since kinds are primarily affected by what laws are accepted by the theory they figure in, it is the process of accepting laws as true which is the only "organic" way of revising kinds, namely the way of operating *within* one theory. (The non-organic way being the theory's rejection and the abandonment of its kinds – and there is a tipping point when organic revisions become abandonments, namely when sufficiently large parts of a theory are changed at once.)

In a nutshell, we cannot revise kinds of one theory in light of the other, since if their kinds don't match up to begin with, then the causal relations won't either. It is the whole theory in which these kinds are defined, along with its whole set of laws, which is to be revised; and it cannot be revised merely in terms of what other theory's causal relations its own line up with. This is not to say that the properties of one theory can't affect the properties of another – only that holding that their kinds or causal relations can be neatly lined up is misleading. The obvious exception being two theories which are purportedly about the same things, and which we thusly assume to make use of the same kinds, but even then, "lining up" kinds or causal relations boils down to finding out which of the two theories is true. Once we find that one theory is superior to the other, salvageable kinds are "revised" in terms of what we now know to be true about them, and non-salvageable kinds are left behind.

These ruminations can seem close to considerations of incommensurability and paradigm change in science (cf. Kuhn 1962 & 2000, Feyerabend 1962). However, this proximity can be misleading, because the reality of theoretical revisions is less dramatic and

more practical than either. The chief practical aspect is, of course, explanatory power. In dealing with potential revisions, our basic assumptions at this point are that we are concerned with two distinct theories which are both explanatorily valuable (i.e. that none of them is plainly wrong) and that one of them is a candidate for revision in light of the other. In other words: The theory up for revision has some explanatory value, but also some categories/kinds which could be somehow improved in light of the other theory. But how can this be if, as just discussed, their meanings, the laws they figure in and the theory which is comprised of these laws cannot be neatly compartmentalised? Well, they can be, if the relevant explanatory relations which are effective for their meaning are respected.

Let's look at an example from two theories which we assume to be interrelated, such as biology and physics: improving our knowledge in physics can improve our biological knowledge. For example, geese are a biological kind due to the fact that there are projectible generalisations such as "geese fly south for the winter". We can assume that there is a neurobiological cause for each individual goose's flying south for the winter, and thusly, that a physical description of this mechanism could improve our knowledge about geese. However, it can't just be *any* physical description of this mechanism, even if it were a description which would turn out to be extremely enlightening for physics. Rather, the physical description needs to be *biologically explanatory*. For instance, if the physical description implies that there are subtle differences in mechanisms which have a systematic bearing on said biological law – such as geese having mechanism *A* flying south slightly earlier than those having mechanism *B* – then we could have reason to introduce *A*-mechanism-geese and *B*-mechanism-geese in biology. In such a case, the relevant biological kind would be respected, since it is acknowledged that the corresponding law is imperative for determining the kind. So, the category is "revised" in the sense that it is differentiated, that it is cut at more precise joints in order to gain an increase in explanatory power. This is how a biological kind can be revised in light of physics (assuming it is in fact the physical description which explains the difference in biologically relevant behaviour).[107] Generally, the idea is that:

[107] However, it should also be clear that any single insight, even if it does pertain to explanatorily relevant characteristics, need not by itself lead to revising the biological theory. Rather, different characteristics need to be balanced, and that geese can produce offspring with one another plausibly outweighs their flying south for the winter. However, if an additional relevant distinction can be introduced to one theory in light of insights from another one, while all other explanatory characteristics stay the same, then no such balance has to be achieved, and revisions can be uncontroversially justified. If other explanatory characteristics do *not* stay the same, balance needs to be achieved, and revision can be more controversial. Whereas, if explanatory characteristics are not respected at all, revision would be nonsensical.

(1) If there is a law F → G

(2) and a systematic relation between F and a kind K from another theory

(3) and K can be analysed as the disjunction "H or I", with H and I being kinds (or exclusively composed of kinds) from the same theory as K

(4a) and "F in terms of H" allows for an explanatory differentiation of G

(4b) or both "F in terms of H" and "F in terms of I" allow for an explanatory differentiation of G

(5a) then (if 4a is true and 4b is not) we have reason to introduce H-type Fs

(5b) or (if 4b is true) to split up Fs into H-type Fs and I-type Fs.

In our example, (1) is the biological law that geese fly south for the winter, (2) specifies that there is a systematic relation between geese and a neural mechanism (namely their typically having it), (3) says that the neural mechanism is found to actually comprise of two distinct neural expressions which, whose specific expression A or B can further specify the biological effect (such as A causes flying south a day earlier than B and B causes flying south a day later than A) (4b), so that we could revise the biological kind "geese" by introducing A-mechanism geese which fly south a day earlier than B-mechanism-geese (5b). Of course, disjunctions (here: F = H or I) could potentially comprise of as many terms as would make an explanatory contribution to the causal effect specified in (1).

If we stick with a pragmatic scientific realism, i.e. with accepting kinds as real insofar as they serve valuable explanatory roles, then the only potential reason for revising kinds is that restating or adjusting one theory in relation to another increases explanatory power. If this ultimate criterion for revising kinds is glossed over confusion can ensue, such as in the form of attempting to cash out explanatory but "unnatural" kinds in one theory against "natural" ones in another (see footnote 106). For example, two views about emotions have recently been clashing in psychology and cognitive neuroscience: In the one corner is *constructionism*, "the view that emotions of a certain kind are constructed out of more general brain structures whose function is not specific to emotions of that kind, or even to emotions at all" (Humeny et al. 2012: 153). In the other we find *locationism*, which holds that the "category *emotion* and individual categories such as *anger, disgust, fear, happiness, sadness* (and perhaps a few others) are respected by the body and brain" (Lindquist et al. 2012: 122). While it is not self-explanatory what it means that emotions are "constructed out of more general brain structures" or that emotion categories are "respected by the body and brain", a weak interpretation of these claims amounts to opposing hypotheses about what kind of neural

structures underlie emotional processing and corresponding behaviour. Such an interpretation is comparably weak because the truth of any of these hypotheses does not entail the need for modifying or abandoning emotion categories. Rather, the results which would make either claim true would serve to specify which neural structures or properties are responsible for the instantiation of states falling under these categories. For example, to say that someone is angry can be true regardless of whether the processing and behavioural control related to this person's angriness is widely distributed, whether it involves several neural mechanisms which are not specific to emotions, and so on. To deny this would mean to conflate emotional and neural categories, in which case there would be no need for intertheoretical revision to begin with (but perhaps for *intra*theoretical revision).

But as a strand of the discussion shows, we are in fact dealing with stronger claims which insinuate that emotion categories clash with more "natural" kinds:

> "In emotion research, measures of emotional behavior are often used to infer the existence of underlying emotion mechanisms. (…) Scowling and crying are taken to be observable evidence that the causal mechanisms for *anger* and *sadness* have been triggered. (…) If emotions are distinct kinds that correspond to real distinctions in nature (i.e., distinctions in the brain and body), then examining the observable outputs for each emotion should give evidence of these distinctions. (...) Questions about the structure of emotion responses (such as the structure of self-report or the structure of facial behaviors) are really questions about whether *anger, sadness, fear*, and so on are the natural kinds that constitute the building blocks of emotional life, and are therefore the most appropriate categories to support scientific induction" (Barrett 2006: 34 f.).

As we have just seen, what is a "real" and "natural" distinction is directly equated with what is a distinct property of the brain and body. It is an unresolved ambiguity that leads to this equation: The natural-kinds-view of emotions is characterised as holding that (certain) emotions are "natural kinds, or phenomena that exist independent of our perception of them. Each emotion is thought to produce coordinated changes in sensory, perceptual, motor, and physiological functions that, when measured, provide evidence of that emotion's existence" (ibid.: 28). The ambiguity lies in the fact that the first claim may well be true even if the second isn't. Phenomena which exist independently of our perception of them need not be characterised in terms of intrinsic properties of an agent, especially if they are of an intentional nature, which usually invokes a relation between an agent and their

environment.[108] I will come back to this ambiguity later. Also note that what isn't at stake is whether emotional categories fulfill their explanatory roles, but whether they group together "projectable [sic!] property clusters"[109] (ibid.: 33):

> "One way to establish the presence of an abstract construct like *anger, fear*, or *sadness* is to demonstrate that each has measurable effects that are highly correlated. From a purely psychometric standpoint, psychologists assume that if measures are highly correlated, then they must derive from a common cause (in this case, the emotion). If measures are weakly correlated, then psychologists typically conclude that the measures have separable causes and do not give evidence of the construct in question. As a result, the extent of correlation between measurable responses provides a psychometric test of whether or not a construct exists. In this case, such correlations provide a way of testing whether or not kinds of emotion exist as definable categories" (ibid.).

Barrett then goes on to report that correlations between measurable effects associated with emotion categories have been empirically tested since the late 1960s, yet these have consistently turned out so weak that "even the strongest correspondences within emotion categories are weaker than those observed for broad affective dimensions" (ibid.). This means that if we were to pick an emotion category and measure correlations between instantiations of its characteristic effects, such as "facial movements, vocal signals, changes in peripheral physiology, voluntary action, and subjective experience" (ibid.), then we would get a lower correlational measure than if we were to correlate "facial behaviors, reports of experience, and peripheral nervous system activity (…) [with] affective properties of valence and intensity" (ibid.). In the study cited by Barrett, "[a]ffective evaluation is defined as the degree to which pictures are judged as (un)pleasant and arousing" (Lang et al. 1993). So, to put it bluntly, there seem to be nicely measurable bodily signature responses to finding a picture "(un)pleasant and arousing", whereas the signature response to being angry pales in comparison.

Does the lack or comparative weakness of emotional signature responses undermine the explanatory credibility of emotion categories? Well, for one, if there are no nicely observable signature responses, then emotions themselves may not be (sufficiently or reliably)

[108] I take the notion that emotions are independent of our perception as replaceable with the notion that emotion-ascriptions potentially express objective facts. That is, it should be nonsensical to say that whether someone is in an emotional state can neither be either true nor false. So, the following debate can be read as being about whether what makes such ascriptions true are intrinsic states of an agent, rather than about whether such ascriptions are vacuous.

[109] The use of this concept evokes Boyd's "Homeostatic Property Cluster Theory" (cf. Barrett 2006: 29, Boyd 1999), which relies on Goodman's notion of "projectible predicates" (Goodman 1983).

observable, and so they run the risk of failing to constitute explanatory objects in reliable empirical theories. Thus, proponents of the view that mental states such as emotions make for kinds in empirical theories might be in serious trouble (compare the enthusiastic defense of the scientific reputability of mental states in Fodor 1989: ch. 1). That mental states must in principle be tied to observable behaviour is a classic view in analytic philosophy and rarely disputed by anyone who believes that theories about mental states have genuine explanatory value (see I.7). As Sellars noted, "the fact that overt behavior *is* evidence for (…) [instantiations of psychological states] *is built into the very logic of these concepts*, just as the fact that observable behavior of gases is evidence for molecular [states] (…) is built into the very logic of molecule talk" (Sellars 1997: 107, §59). I will call this notion that explanatory psychological theories need to be (or are always) systematically tied to potentially observable behaviour the *observability constraint*.

The categories used by said empirical psychological theories, such as anger, are vindicated if laws like "ceteris paribus, A is more likely to share something valuable to her with B when not being angry at B than when she is" have explanatory value. In other words, that anger is an explanatory kind in any one such theory means that "anger" is a technical term defined by the laws in which it is invoked. Thus, the meaning of anger is extensionally defined using the totality of laws invoking anger (compare Lewis 1972: 204, 207 f.).[110] In our example, it would work in the following way: Given only said law, "anger" is defined as any directed property, which, whenever A has it and it is directed at B, makes A less likely to share something valuable to her with B. If we only had this one law at our disposal, then a host of other psychological properties (such as jealousy, hatred, and so on) or perhaps physiological conditions would fulfill this definition. So, another constraint has to be introduced which specifically singles out anger as the property invoked by said law, namely that there is a host of laws which, when taken together, single out anger instead of jealousy, hatred, or a physiological condition. (So, for the definition of the kind "anger", we would

[110] Which does not mean that there is a fixed totality of such laws, only that, at any given point in time, there is a number of such laws extensionally defining the concept (many of which plausibly endure over time). Neither does it mean that, in order to properly apply the concept, we need to have learned all the laws characterising it; but we certainly need to have learned a sufficient amount. That is, in order for someone to apply a specific concept, they need to have learned enough laws invoking it to be able to make a distinction between the applied concept and other concepts (this wouldn't involve too many laws if all the concepts we would ever apply were, say, anger and joy, but with each additional concept, we need to learn additional distinguishing laws). The view that we learn such concepts by witnessing stereotypical instantiations or expressions of them – for example: we learn what anger is by witnessing enough angry people doing things out of angriness, and extrapolating the common property between all these events – would mean that we extrapolate mental laws from their instances by hypothesizing that the paradigmatic instantiations we've witnessed are paradigmatic *because* they are instantiations of the laws governing the mental concepts.

arrive at a conjunction of the roles it plays in anger-laws.)[111] I will call this the *specificity constraint*.

I do not claim that these constraints are exhaustive for what makes a good theoretical kind. For instance, they do not say anything about whether the theory's laws are explanatorily valuable, which we would need to assume (and which I in fact have assumed earlier) in order for determining whether the role theoretical kinds play is one serving explanation. Rather, I have singled out these two constraints because, on the one hand, it is the observability constraint which seems to be under direct attack by the aforementioned studies, and, on the other, the specificity constraint is the one which tells us how to define emotional kinds and what role signature responses might play for defining them. So, we should consider in turn whether the two constraints are touched by the cited results, namely that measurable bodily responses generally correlate more strongly with affective valence and intensity than with emotion categories. (In the following, I will set any methodological criticism aside and treat these results as facts.)

So, what about the observability constraint? At the very least, the studies do not conclusively show that it is violated. What research tells us is that bodily responses may not yield sufficient criteria for emotion ascriptions, but it says nothing about the lack of contextual information. As mentioned, intentionally characterised psychological states typically take relations between an agent and their environment into account, rather than just the agent's intrinsic properties (compare I.8.3 – I.8.5). For example, if I shove someone, they may not show any behavioural sign of anger; yet, my attributing anger to them can be justified on the grounds that I shoved them and that shoving generally angers people. Similarly, external circumstances are typically marked as providing justifiable evidence for emotional states: we often call events "joyful", "frustrating", "sad", "annoying", etc., which means that being in any one such context can be justification enough for attributing the associated emotion (at least with a certain likelihood). So, that bodily expressions or responses fail to generally make for definite criteria does not imply that we generally lack criteria for attributing emotions. Rather, bodily responses make for criterial evidence (cf. Dennett 2007: 74) which is both supplementable as well as defeasible by contextual information. So, it is readily acceptable that an intense somatic state correlates more strongly with, say, a kind of facial behaviour than anger or sadness generally do.

[111] It may be historically true that purported natural kinds are first lumped together in virtue of their surface features, rather than in terms of their theoretical explanatory value (so that, perhaps, anger is thought of as the category which lumps together states producing certain facial expressions), and that only through ongoing scientific investigation the groupings are revised in terms of underlying causal features (cf. Craver 2009, Reid 2002), which would enable us to come up with said extensional definition of the meaning of a theoretical kind.

Do the cited results touch the specificity constraint? Only if it can be shown that the states which compete with emotions can take their role in such laws as "ceteris paribus, A is more likely to share something valuable to her with B when not being angry at B than when she is". While it cannot be ruled out that there are such somatic states, the research establishing correlations between facial behaviors, reports of experience, peripheral nervous system activity and affective properties of valence and intensity clearly does not even begin to suggest this. (And neither does it show that laws such as the one just cited are in any way moot.) Trivially, the disjunction between all somatic states underlying emotion would qualify as subplanting emotions, but surely it would be hypocritical to claim that this set is purely governed by theories about somatic states, since it clearly owes its existence to its picking out states governed by emotion-ascriptions (cf. Fodor 1974, see also II.8.4.5).

Let's take a closer look at how neural kinds play into the claims marking the dispute between locationism and constructionism. For one, what exactly are these "neatly delineated" neural kinds associated with locationist claims? Obviously, we are not just talking about individual neurons and their properties, but about functionally individuated networks of neurons, brain areas or neural mechanisms:

> "All natural kind models share the assumption that different emotion categories have their roots in distinct mechanisms in the brain and body. The mechanisms underlying discrete emotion categories have been discussed as residing within particular gross anatomical locations (...) or networks (...) in the brain. These models constitute a locationist account of emotion because they hypothesise that all mental states belonging to the same emotion category (e.g., fear) are produced by activity that is consistently and specifically associated with an architecturally defined brain locale (...) or anatomically defined networks of locales that are inherited and shared with other mammalian species" (Lindquist et al. 2012: 122 f.).

Networks and brain areas can be defined by their architecture (i.e. by spatial or structural properties) or physiognomically (i.e. by their location and structure relative to the organism's body), but also functionally (in terms of an array of neurons connected in virtue of a certain task they're dedicated to performing). On the other hand, mechanisms are primarily defined functionally (cf. Sullivan forthcoming): it is only once an array of neurons has been found to constitute a mechanism that this array can secondarily be described in terms of its architecture or physiognomy. Mixing these descriptions can cause confusion since functional descriptions aren't innocently "natural" (i.e. not primarily physical), and even architecture or physignomy needn't be once it depends on mechanistic individuation.

Yet, as we have seen in the passage just cited, even mechanistically individuated neural kinds are commonly labelled as "natural kinds" and uncritically opposed to categories such as emotions, which are primarily functional categories. This opposition is artificial insofar as the functions which are the basis for individuating emotions are also the basis for identifying neural mechanisms, and it is misleading insofar as neural structures are not only treated as implementations of such functions (cf. Nathan & Del Pinal forthcoming: 5.3), but as *the* relevant kinds in theories explaining phenomena connected to emotional processing and behaviour. Now, either they are such relevant kinds because they are conceived of as mechanisms, i.e. as connected to the functions underlying emotions; then these kinds are explanatorily *connected* rather than opposed. Or they are defined in non-functional terms (such as physical kinds, or spatial [non-functional] architecture or physiognomy), in which case their explanatory role is at best indirectly connected to emotional categories. "It would be very surprising indeed if the brain were organised into spatially discrete units that conform to our abstract categorisations of behavior" (Valenstein 1973: 142 f.), but thankfully, being organised into spatially discrete units has no direct explanatory bearing on our "abstract" categorisations of behaviour to begin with.

Consequently, the claim that there are "distinct mechanisms in the brain and body" underlying emotions is misleading as well. Mechanisms are distinct if they are explanatorily distinct, and there is no direct relation between spatial distribution and explanatory role. If they are mechanisms pertaining to emotional categories, then their being distinct mechanisms means that they are distinct regarding explanations pertaining to phenomena related to emotions. Typically, evidence for whether an emotion category is instantiated is different from evidence about whether a neural category is instantiated. So, just as in the example of physical mechanisms potentially explaining why some geese fly south earlier than others, neural properties make a difference to emotion-categories if they are explanatorily relevant. For example, if (purely hypothetically) a widely distributed neural state were to result in more severe aggressive anger-related behaviour than a densely located one, then this fact would influence angriness-categories. But finding out whether locationism or constructionism is true regarding one emotional state need by itself have no consequence for the respective emotional category (even though this insight may constitute a considerable leap forward for the neuroscience of emotion). So, a mental-neural mapping which respects explanatory categories can allow for a mental state to be implemented in all manner of disjunctive ways, as long as the properties of the associated neural state do not make a difference to the explanatory value of the mental state (see II.8.4.4). To be sure, it may constitute one indirectly: Perhaps some

distributions end up causing methodological problems, namely if not all relevant brain areas are or can be "read" by a given method, or if distributed data is harder to read than centrally localised data, or if relevant activation patterns are just so distributed as to exceed limits of, say, fMRI resolution (see II.4).

Conversely, one and the same brain area or pattern of activation can be involved in processing pertaining to diverse mental states. For instance, "[p]sychopaths, in whom striking emotional anomalies are strongly correlated with specific brain anomalies, appear to challenge constructionism" (Humeny et al. 2012.: 153) – i.e. one psychiatric category is associated with localised brain damage –, yet "[a]ggression appears to be how psychopaths respond to a wide variety of situations about which they have a wide variety of feelings. If this is true, aggression in them is likely to be associated with a wide range of emotions, not just anger. This conclusion supports constructionism" (ibid.: 154). Similarly, Lindquist et al. note:

> "Overall, we found little evidence that discrete emotion categories can be consistently and specifically localised to distinct brain regions. Instead, we found evidence that is consistent with a psychological constructionist approach to the mind: A set of interacting brain regions commonly involved in basic psychological operations of both an emotional and non-emotional nature are active during emotion experience and perception across a range of discrete emotion categories" (Lindquist et al. 2012: 121).

Once again, we find that our earlier distinction between spatial and mechanistic characterisation applies: Since emotional categories are directly related only to mechanistic explanation, and mechanisms are not (merely) characterised in terms of spatial structure or anatomy, the mere fact that a brain region which is spatially, structurally or anatomically characterised *as being a distinct brain region* (i.e. as distinguishable from other spatially, structurally or anatomically characterised regions) underlies several emotions has no direct bearing on emotional categories. (We should not exclude that it has an indirect bearing, but the cited research gives us no conclusive reason to assume that it does, since it only argues for neural categories having a direct bearing on emotional categories.) To say that several different mechanisms are executed in the same region is no inherent contradiction, and neither is it to say that the execution of one mechanism is spatially or anatomically distributed.

As we have seen, intertheoretical revisions, construed as revisions of theoretical kinds of one explanatorily valuable theory in light of another, can work under the constraint that the explanatory power of the revised theory is respected. This is because the meaning of a

theoretical kind is defined in terms of how it contributes to the explanatory power of the theory it is evoked by. Thus, if we treat two theories as distinct – as employing different explanatory strategies, as stipulating different causal groupings, as having different scopes – then there is no reason to assume that the kinds or causes evoked by different theories neatly "line up" or non-disjunctively map onto each other. (In fact, their divergence in meaning may make it difficult for us to judge whether they do.) Making them want to line up forcibly, namely only in explanatory terms of the revising theory, but not of the theory under revision, has no constructive effect at best, or at worst disrupts the revised theory.

As our review of experimental results regarding emotional signature responses showed, behaviourally observable or measurable somatic states can underdetermine emotion categories (cf. Barrett 2012: 421). In other words, the differences we routinely make between such categories do not necessarily imply a difference in somatic or affective state, or in neural states producing the respective somatic or affective states. As Barrett puts it, "perhaps one of the most important questions that remains is why perception-based judgments routinely produce evidence in support of emotion categories, even as instrument-based measurements do not" (Barrett 2006: 49). Following Barrett, I suggest that we should assume that contextual information beyond the somatic or behavioural state of a person contributes to the mental ascription (compare Barrett 2012 for her view on the "social reality" of emotions). This view that intentionally characterised psychological states do not merely reflect intrinsic facts about an agent, but also extrinsic ones such as her relation to the environment, meshes exceedingly well with the form of anti-individualism adopted earlier (see I.8 and Waskan 2006: 89 f.). Experimental results are only inconsistent with the natural kind view of emotions (cf. Barrett 2006: 49) as long as these natural kinds are construed as referring to an agent's intrinsic states. However, if we assume that common emotion categories are anti-individualistic, the experimental results do not constitute grounds for doubting the explanatory value of these categories. We would in fact *lose* explanatory power if we were to revise emotion categories *only* in terms of measurable bodily states, since the bodily responses underlying the emotional one need not single out the respective emotional state.

As discussed in II.8.4.4, that a mental term translates to a disjunctive neural one, or vice versa, by itself constitutes no obstacle for a translation; it is only when such disjunctive terms cannot be correlated significantly to an explanatory term in the other theory – and cannot be distinguished from contradicting terms – that problems arise. Even though the studies under review suggest some mental-neural cross-cutting, and we should therefore expect translated terms to be disjunctive, this does not establish that translational efforts

would be impeded by way of catastrophic ambiguity. As Valenstein pointed out, it would be very surprising indeed if intentional categories would coincide neatly with neural categories (cf. Valenstein 1973: 142 f.). Since intentional and neural kinds are characterised differently, their identity criteria differ substantially to begin with (some may not even have spatial or temporal identity criteria, cf. Nachev & Hacker 2014), and so I suggest we should waive such demands. Neat mappings might happen, but it makes no catastrophic difference to either translated theory if they don't.

II.9. Summary

Representations are prevalent in cognitive neuroscience (II.1). Yet, neurobiology alone does not supply representational concepts (II.2), and its explanatory value does not depend on matters of semantics (II.3). While we can find neural representations to be related to covariance and cause, neither notion supplies sufficient justification for calling neural kinds or structures representational (II.3 & II.4). Instead, we need to think of cognitive mechanisms as neural structures whose joints are carved along their functional roles instead of along their synapses. This functionality supplies teleological descriptions, thus specifying the aim which is needed for characterising the mechanism's directedness (II.5 & II.6).

When describing the in- and output of such mechanisms semantically, the content of "neural representations", which means the neural substrate of a mechanism, can be identified in several ways (II.7). Firstly, it can be identified by determining whether one such mechanism has been evolutionarily selected to fulfill its function (II.7.1), which is to be sensitive to certain environmental cues – cues elicited by the mechanism's intentional object – and to produce a certain functional output (e.g. behaviour which can be criterial for the ascription of semantic content). An evolutionary analysis of a cognitive mechanism consists of two steps: In describing the mechanism's functionality in terms of adaptiveness within a certain environment, and in describing the cause for the mechanism's existence. Secondly, using a dynamic systems approach, other organismic purposes can be singled out as specifying the mechanism's content, chiefly those contributing to homoeostasis (II.7.2). This way of ascribing content does not suffer from evolutionary theory's problem of historicity, which means content can be ascribed to agents independently of facts about their ancestors, and so they can be tested independently of evolutionary methodology. However, there is a

trade-off for solving this problem: Some evolutionarily functional mechanisms cannot be grasped as such when viewed purely through the lens of homoeostasis.

However, evolution does not provide the only possible causes of the formation of such mechanisms, and neither is homoeostasis their only possible purpose. That is, while we generally require the structure of such mechanisms to be established by external requirements, these requirements need not be of an evolutionary nature or pertain to homoeostasis. Aside from practical requirements depending on a specific environment, which have shaped some mechanisms by way of organismic evolution, there can also be social or cultural requirements (II.7.3). These requirements are properly described as norms, conforming to which an organism can learn. Norms are environmental matters of fact: A norm is an external cause which – by way of social mediation – influences the ontogenetic aspects of an organism's neural development. Unlike evolutionary requirements, norms are conventionally established and enforced under agential description. In other words: While evolutionary requirements are contingent on environments, social or cultural norms are contingent on (individual or collective) decisions.

Generally, representations are relational structures holding between intrinsic or organismic and environmental properties connected by way of said teleological principles (II.7.4). Theories using rich content to characterise their explanatory kinds depend on there being norms in light of which to interpret some states as having such content (II.8.1). These norms are not (and are not inferrable from) facts about things which are narrowly characterised, such as neural states. Rather, they are external environmental matters of fact which shape neural development. As such, neural explanations cannot pertain to intentional states without ackowledging such norms as causal factors. Since the behaviour that counts as abiding by norms must be learnable and observable, the conditions under which brain states imply intentional states are specifiable (II.8.2). When relating intentional and neural descriptions, we can construe these norms as being imposed on certain physical properties, specifying whether and how these are interpretable intentionally. We effectively treat such physical properties as signifiers. Yet, the (physical, chemical, neural) processes which cause the instantiation of such properties need not be interpretable intentionally themselves (II.8.3).

If normative criteria make for a constant or specifiable variable, we can use correlations between neural and intentional states as a method for mapping one to another (II.8.4.1). By gathering such correlations holistically we can build a translation manual between neural and intentional states (II.8.4.2). This translation is non-reductive, insofar as it depends on taking teleological principles into account, without such principles being reduced

to the neural state description (II.8.4.3). These principles can themselves amount to sparse explanation, such as in biology, or to rich explanation, such as in intentional psychology. Non-reductivity is primarily due to the need for maintaining these non-physical principles, but in the case of intentional explanation it secondarily follow from the latter's interpretational nature, since interpretation needs to methodologically maximize an agent's rationality. In any case, neuroscientific investigation alone cannot settle the matter whether a specific state or action is directed at a specific object, even though neural processes are constrained and shaped by such norms of directedness.

When mental and neural states do not map neatly onto one another it has been claimed that mental theories can be subject to a revision of their kinds by "lining them up" with causes from behavioural and/or neural theories (II.8.4.5). However, since a theory's kinds and causes cannot be defined independently, kinds can only be modified if it increases their explanatory value relative to the theory which defines their meaning. Also, the notion that it is merely intrinsic properties which single out "natural kinds" underlying emotions mistakenly disregards that they are characteristically individuated anti-individualistically. Further, intentional kinds cross-cutting neural kinds does not directly make any explanatory difference to either intentional or neural theories and therefore does not imply the need for their revision. That is, the mere fact that the neural translation of an intentional state description is disjunctive does not speak against our describing the intentional state the way we do (and vice versa), and, insofar as disjunctive terms do not imply translational ambiguity, it does not impede translation (II.8.4.4).

Conclusion

Understanding intentionality goes a long way toward understanding the mind. Mental states are characterised representationally, and they are related to the world and to one another normatively. In this book I have explored and endorsed the view that the way mental states come to be intentional, and what it means that they are, can be best understood by investigating the explanatory value intentional states have. Assigning meaningful states to us explains our behaviour in ways which the assignation of non-semantic states cannot. This is because both our biologically as well as our socially determined behaviour is regulated by the objects of our mental representations.

Granted, mental states need not all be intentional, and even those that are can still have non-intentional properties. Especially the widely shared view that mental states have a subjective experiential quality to them need not invoke matters of intentionality. Yet, the way these experiences are characterised and referred to – namely as being experiences *of* something – is steeped in intentional vocabulary. For this reason, they need to be connectable to the way we intersubjectively assign meaning: While having experiential states can be private, their being intentional cannot.

Such an assignation consists in systematically connecting observable properties with potentially non-observable "meanings". Here, the observable properties act as signifiers, and while these are grouped together by the meaning they are endowed with, they can be described non-semantically. For example, while each token of the letter *A* has a meaning, namely "A", its material aspects, such as the range of angles between the three characteristic lines, its overall orientation and spatial relation to other signifiers, etc., are describable non-semantically. This is how non-semantic stimuli can elicit semantic mental representations and how individuals can learn all the different tokenings of writing letters, eating non-toxic food, voting, giving promises, and so forth.

While having a mental state can also explain having further mental states – such as feeling treated unjustly can explain being resentful –, they are ultimately invoked to explain actions. Therefore, they are only fully ascribable to agents (or, when used in attenuated form, to sufficiently agent-like entities). The paradigmatic explanatory model used by intentional psychology consists in relating actions to the reasons which elicit them, typically by assigning

beliefs rationally related to the intended action and desires whose content is the action's projected consequence. Other psychological attitudes might work a bit differently, but all potentially play an analogous role in action explanations.

I have argued for a broad content view of mental ascriptions. Broad descriptions are those which take environmental facts into account. According to this view, the ascription of a mental state is best understood as a hypothesis about (A) the obtaining of a certain cognitive structure of the agent (namely, that the ascribed rational relations are causally active in the agent's mind) and (B) the obtaining of certain causal and rational relations between this cognitive structure and external circumstances. For example, having learned to read is the cause for the obtaining of a certain neural structure implementing a cognitive mechanism, which, if it consistently yields a proper output based on a matching input, can be behaviourally identified as executing the function *reading*. And, insofar it has led to the development of a consistently functioning mechanism, worms being nourishing for toads is the cause for the toad's cognitively representing worm-features. So, any such ascription at least implicitly specifies (B) as the cause for (A). In this sense, mental state ascriptions are not reducible to internal ("narrow") state descriptions of the respective agent. Rather, agential descriptions depend on both narrow and broad facts.

In intentional explanation, rational relations among mental states, as well as between mental states and their objects, are methodologically employed to assign mental content. In other words, agents are those who exhibit at least a minimum of consistency, both internally as well as externally. Internal consistency is measured by the logical consistency between mental states as well as the practical consistency between mental states and actions. External consistency is measured by the diversity of behaviourial responses elicited across different instantions of relevantly identical environments under relevantly identical mental states: the higher the diversity, the lower external consistency. For example, if my desire to eat oranges rather than other fruit stays constant, and other factors remain equal, my choice of food should not differ wildly across several instances of being presented assortments of fruit including oranges.

It is sometimes assumed that human beings are too irrational to consistently stick to such psychological laws; but even their straying from them is typically explained by evoking conflicting mental states. We may occasionally seem irrational because not all our motivations are transparent (perhaps not even to us), but hidden motivations are still motivations after all. Crucially, the assumption that agents are largely rational is not an empirical hypothesis about agents, but a condition for viewing them as agents in the first

place. If fundamental conditions of rationality turn out not to apply to someone, then we cannot describe them as an agent, as performing actions, or as having mental states. Thus, while intentional psychology is concerned with an individual's cognitive structure, it also transcends it by invoking rational explanation and justification: What is rational cannot be deduced from individual psychological constitution, but a certain psychological constitution is necessary to act rationally.

I have also argued for intentional realism: That the practice of ascribing mental states is generally justified because psychological laws pick out explanatory relations between kinds of cognitive states, kinds of behaviour, and kinds of objective circumstances. Each instance of a mental state ascription is justified if it adheres to criteria specifying under which objective set of available evidence the ascribed mental state obtains or under which it is its (rational, functional or statistical) cause or effect. One such ascription is true if the cognitive structures or causal relations required, assumed or implied by the mental state ascriptions actually obtain.

Said psychological laws are reliably applicable because human beings are typically cognitively able to systematically associate bearers of meaning with environmental properties, conditions or events. Meanings capture such properties, conditions or events which we can have a psychological attitude towards (i.e. an intentional mental state). At the same time, meanings are the operative aspects of interpretation, namely the objects whose ascription to someone makes their actions interpretable, i.e. largely reasonable.

The *ascription* of mental content depends on knowing the applicable psychological laws, so they depend on theoretical knowledge. *Having* mental content can, but need not, depend on such theoretical knowledge. I have marked this distinction by calling the kind of content which depends on such knowledge "rich" and the content which does not "sparse". Typically, sparse representation can be explained in biological terms, while rich representation relies on social learning. For instance, learning social norms is not a precondition for being able to be hungry. While potential objects of the agent's hunger can be specified as the intentional object of hunger, the agent's attitude and the resulting directedness of her behaviour need not be explained by social norms (it *can* be when hunger is socially conditioned) but can be explicated in terms of organismic/biological directedness. However, an agent can also hunger for cheddar cheese, intend to play chess every Sunday, or know that she should have sold her stocks before the last stock market crash. These mental states are rich, insofar as they depend on the agent having been instructed about certain theoretical and symbolically mediated classifications in order for these to become potential objects of her

mental states. In other words, the fact that some psychological laws apply to the agent is necessarily owed to causal properties of representations themselves (although not sufficiently, since the agent's sparse – narrowly characterised – cognitive makeup contributes to her internalizing such effects).

While I have only mentioned them in passing, there are two viable arguments to be made to refute intentional antirealism, i.e. the view that mental states are explanatorily inert, illusory and/or unscientific. A minimal argument says that mental states explain a wide range of everyday human actions (such as why someone shows up where she promised to be), and that at least at this time, there is no explanation in any other science which comes close to this explanatory power in regard to such phenomena. A less boastful argument holds that even if there were, say, neurological accounts which would predict and explain why someone did in fact show up where she promised to (and which could help us construe giving this promise in neural terms as well), the explanation would still have to recur to things beyond neural matters: namely the content of the promise and the conventional rules underlying the practice of promise-giving. In this sense, we would not abandon the notion of intentionality, but rather put intentional psychology on more solid footing.

When we use representations in neurobiology to explain behaviour, we typically do so sparsely, by relating bodily facts about an agent (from sensory input via nervous/neural processing, which includes electrical and chemical transduction as well as storage of neural "information", i.e. changes in neural structures to accommodate and influence future transduction based on past transduction, to innervation of muscles) to the biological function and/or heritage of the respective neural mechanism. The function specifies the directedness of the "representation", i.e. of those internal structures which we can single out as a mechanism.

Beyond such sparse representations, we can invoke neurobiological methods to explain psychological properties invoked by intentional psychology. To relate neurobiology and intentional psychology, intentional states are also construed as functional, i.e. as characterised by input-output-pairs which are related by neural processing. What matters for intentional psychology is that either input or output or both are interpretable: that they are signifiers. Neurobiology can then supply a nomological ("lawlike") description for how the output's material aspect is derivable from the input's material aspect via neural processing.

A "law" in the natural sciences is a kind of statement which justifies generalisation about its content based on inductively collected evidence. For example, a law governing the release of a neurotransmitter under certain conditions justifies us to expect this release under all instantiations of these (or similar enough) conditions. The explanatory concepts used in

nomological explanation are kinds (i.e. projectible/generalisable properties) and causes (relating such generalisable properties). Mechanistic explanation, which we are prone to find in cognitive neuroscience, uses lawlike explanation but adds explanatory concepts beyond kinds, causes and laws, such as temporal and spatial relations. Laws typically explain by way of integration into higher-order laws. For example, causal laws about the release of neurotransmitters should be integratable into higher-order physicochemical laws.

While neurobiological states are described narrowly and non-representationally, correlations between intentional and neurobiological states can contribute to a translation between the terms denoting such states. What delineates the scope of any such translation is the extent to which distinctions made in narrow (neurobiological) descriptions can support the respective inferences to broad (intentional) concepts. One example for an endeavour to come up with algorithms for inferring broad from narrow states is the development of "brain decoders": Recent studies have shown that computers can be "trained" to output semantic content when fed activational information about certain brain regions.

Apart from such experimental studies we also have theoretical reason to believe that the respective inferences should be viable. For example, it is apparent that, say, the English language can be learned and spoken, and that, to some degree, speakers need to rely on stable neural processing in order to learn it and to be able to keep on speaking it. While our behavioural observations can of course not determine whether the respective processing is stable in narrow terms (i.e. in terms of sustained neural structures), it does determine its stability in broad terms: Namely in terms of significantly outputting the correct signifiers depending on matching input. Given that each behavioural input- and output-signifier is non-representationally describable and that their matching is caused neurally, neurobiological processing and narrow descriptions of such signifiers *must* conflate: Neural processing operates on material input and yields material output, and either or both are signifier-tokens for associated representational content. While the association between material tokens and meaning still depends on broad facts, namely a fitting environment, it is reasonable to assume that many biologically or socially relevant conditions under which narrow signifiers (such as the perception of elongate objects moving parallelly to their longitudinal axis and the string of letters making up the word "rabbit") consistently map onto stable meanings (such as worms and rabbits). So, it is exceedingly plausible that facts about neural structures carrying out the respective processing can track such meanings, that the latters' in- and/or outputs are consistent with external matters of fact specifying the meaning of signifiers, and that neurobiological knowledge can be exploited to build said translation function. To assume the

alternative, namely that neural in- and outputs merely singularly coincide with semantically interpretable states but not systematically or consistently so, would leave our command of matters of semantics as well as some of the functional stability of our behaviour utterly mysterious. Therefore, we should expect ongoing research into the physicochemical basis of neural processing, as well as into neural correlates of semantic properties, to contribute to a translation between intentionally and neurobiologically characterised states.

This translation potentially serves interests going beyond our common everyday means of intentionally describing and explaining agents. For one, tracking intentional states non-behaviourally is the standard aim driving the development of brain-decoders: to supply agents who have lost the ability to express themselves behaviourally with new kinds of means to interact with the world, and, more controversially, to extract evidence for the ascription of intentional states from uncooperative agents who are deemed (potentially) criminal or harmful. Secondly, neurobiological insights may indeed yield finer intentional categories, insofar as systematic associations between intentional causes and intentional consequences might be trackable more concisely than with our classic behavioural methodology. And, since tracking such relations is one domain of intentional psychology, these insights can lead to theoretical refinements and/or revisions.

Of course, neurobiological insights can't directly guide the *normative* associations between such in- and outputs; but even there these insights can have an impact, insofar as, if we accept that a normative "ought" needs to imply a psychological "can", we should not expect more from agents than they are physically able to deliver. Ever since its inception, psychology has been uncovering limits of agency, leading to our retracting ascriptions of responsibility in specific cases. For example, if we find that certain contexts are prone to yield biased behaviour, this behaviour is not ascribable as agential tout court. We should expect neurobiological insights to add to such retractions (and perhaps even expansions) of agential responsibility.

Yet, there is no question that the largest revenue of an increasingly concise ascription of intentional states lies in more properly situating agents in the world, in taking stock of the relations between them and their surroundings, their environmental niche, their community, their culture. The ways in which we scientifically do so can be manifold – they need neither rely on any specific symbolic form, nor on any special scientific methodology. They simply need to capture these relations in order to explain why we do what we do.

Bibliography

- Adam, C. & Tannery, P., eds. (1964–1974): *Oeuvres de Descartes*, 13 vols.. Paris: Vrin/CNRS.
- Anderson, J. (2009): *Cognitive psychology and its implications*, 7th ed.. New York, NY.: W. H. Freeman/Times Books/ Henry Holt & Co.
- Anderson, S.W., Bechara, A., Damasio, H., Tranel, D. & Damasio, A.R. (1999): "Impairment of social and moral behaviour related to early damage in human prefrontal cortex", in: *Nature Neuroscience* 2: 1032–1037.
- Aquinas, St. Thomas (1272/1952): "Treatise on Man", in: *Summa Theologica*, transl. by Fathers of the English Dominican Province, revised by D. Sullivan, published by W. Benton, Volume 19 in the Great Books Series. Chicago: Encyclopedia Britannica, Inc..
- Armstrong, D. M. (1981): "What is consciousness?", in: Armstrong, D. M., ed., *The Nature of Mind*. Ithaca, NY: Cornell University Press: 55–67.
- Ashby, R. (1954): *Design for a Brain*. London: Chapman & Hall LTD.
- Athanasopoulos, P., Bylund, E., Montero-Melis, G., Damjanovic, L., Schartner, A., Kibbe, A., Riches, N., Thierry, G. (2015): "Two Languages, Two Minds: Flexible Cognitive Processing Driven by Language of Operation", in: *Psychological Science* 26 (4): 518-526.
- Baron-Cohen, S., Leslie, A. and Frith, U. (1985): "Does the autistic child have a 'theory of mind'?", in: *Cognition* 21: 37-46.
- —— (1986): "Mechanical, behavioral, and intentional understanding of picture stories in autistic children", in: *British Journal of Developmental Psychology* 4: 113-125.
- Barrett, H. & Kurzban, R. (2006): "Modularity in Cognition: Framing the Debate", in: *Psychological Review* 113 (3): 628–647.
- Barrett, L. (2006): "Are Emotions Natural Kinds?", in: *Perspectives on Psychological Science* 1: 28-58.
- —— (2012): "Emotions are real", in: *Emotion* 12: 413-429.
- Bechtel, W. (1998): "Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge in Cognitive Science", in: *Cognitive Science* 22 (3): 295-318.

- Bechtel, W. & Wright, C. (2009): "What is psychological explanation?", in: P. Calvo and J. Symons, eds., *Routledge companion to philosophy of psychology*. London: Routledge: 113-130.
- Beck, H. (2014): *Hirnrissig.* Munich: Carl Hanser.
- Beer, R.D. (2000): "Dynamical approaches to cognitive science", in: *Trends in Cognitive Sciences* 4: 91–99.
- Bennett, M. & Hacker, P. (2003): *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.
- Bhat, P. R. & Sahu, G. (1998): "Quine On Observation Sentences", in: *Indian Philosophical Quarterly* 25 (3): 403-418.
- Bischof, N. and Zehetleitner, M. (2015): *Struktur und Bedeutung*, 3rd ed. Bern: Verlag Hans Huber. [forthcoming]
- Block, N. (1991): "What Narrow Content is Not", in: Loewer, B. and Rey, G., eds., *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell: 33-64.
- —— (1995a): "On a confusion about the function of consciousness", in: *Behavioural and Brain Sciences* 18: 227-47.
- —— (1995b): "The Mind as the Software of the Brain", in: *An Invitation to Cognitive Science Vol.3: Thinking*, 2nd ed.. Cambridge, MA.: The MIT Press: 170-185.
- —— (2003): "Searle's Arguments Against Cognitive Science", in: Preston and Bishop, eds., *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford: Oxford University Press: 70-79.
- —— (2007): "Consciousness, accessibility, and the mesh between psychology and neuroscience", in: *Behavioral and Brain Sciences* 30: 481-548.
- Boehlich, W. (1990): *The Letters Of Sigmund Freud To Eduard Silberstein 1871-1881,* transl. by A.J. Pomerans. Cambridge Massachusetts: The Belknap Press of Harvard University Press.
- Botvinick M, & Cohen, J. (1998): "Rubber hands 'feel' touch that eyes see", in: *Nature 391* (6669): 756.
- Boyd, R. (1999): "Kinds, Complexity and Multiple Realization: Comments on Millikan's 'Historical Kinds and the Special Sciences'", in: *Philosophical Studies* 95: 67-98.
- Braillard, P. & Malaterre, C. (2015): *Explanation in Biology. An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences*. Dordrecht: Springer.

- Brandom, R. (1994): *Making it explicit. Reasoning, Representing, and Discursive Commitment*. Cambridge, MA.: Harvard University Press.
- —— (2002a): *Tales of the Mighty Dead – Historical Essays in the Metaphysics of Intentionality*. Cambridge, MA.: Harvard University Press.
- —— (2002b): "Non-Inferential Knowledge, Perceptual Experience, and Secondary Qualities. Placing McDowell's empiricism", in N. Smith, ed., *Reading McDowell. On Mind and World*. London and New York: Routledge.
- Brentano, F. (1973): *Psychology from an Empirical Standpoint,* transl. by A.C. Rancurello, D.B. Terrell and L. McAlister. London: Routledge.
- —— (1995): *Descriptive Psychology,* transl. by Benito Müller. London: Routledge.
- Brooks, R. (1991): "Intelligence without representation", in: *Artificial Intelligence* 47: 139–159.
- Brosnan, Sarah F. & de Waal, F. (2003): "Monkeys reject unequal pay", in: *Nature* 425 (6955): 297 – 299.
- Brower, J. & Brower-Toland, S. (2008): "Aquinas on Mental Representation: Intentionality and Concepts", in: *The Philosophical Review* 117: 193-243.
- Buller, D. (2006): "Evolutionary Psychology: A Critique", in: E. Sober, ed., *Conceptual Issues in Evolutionary Biology*, 3rd edition. Cambridge, MA.: MIT Press: 197–214.
- Burge, T. (1979): "Individualism and the mental", in P. French, T. Uehling Jr., and H. Wettstein, eds., *Midwest Studies in Philosophy Vol. 4 (Metaphysics)*. Minneapolis: University of Minnesota Press.
- —— (1986): "Individualism and Psychology", in: *Philosophical Review* 95: 3–45.
- —— (2007): *Foundations of Mind: Philosophical Essays, Volume 2*. Oxford: Oxford University Press.
- Buss, D. M. (1995): "Evolutionary psychology: A new paradigm for psychological science", in: *Psychological Inquiry* 6: 1–30.
- Carnap, R. (1931): "Die physikalische Sprache als Universalsprache der Wissenschaft", in: *Erkenntnis*, Vol. 2: 432–465.
- Carroll, J., ed. (2004): *Readings on Laws of Nature*. Pittsburgh: Pittsburgh University Press.
- Carruthers, P. & Smith, P. (1996): *Theories of theories of mind.* Cambridge: Cambridge University Press.

- Carruthers, P., Laurence, S. & Stich, S., eds. (2008): *The Innate Mind, Vol. III, Foundations and the Future*. New York, NY.: Oxford University Press.
- Chakravartty, A. (2013): "On the Prospects of Naturalised Metaphysics", in: Ross, D., Ladyman, J. & Kincaid, H., eds., *Scientific Metaphysics*. Oxford: Oxford University Press: 27-50.
- Chalmers, D. (1996): "Does a Rock Implement Every Finite-State Automaton?", in: *Synthese* 108: 309–33.
- —— (2010): *The Character of Consciousness*. New York, NY.: Oxford University Press.
- Chalmers, D. & Clark, A. (1998): "The Extended Mind", in: *Analysis* 58 (1): 7-19.
- Chisholm, R. M. (1958): "Sentences about believing", in: H. Feigl, M. Scriven, and G. Maxell, eds., *Minnesota Studies in the Philosophy of Science,* Vol. 2. Minneapolis: University of Minnesota Press: 510–520.
- Cholbi, M. (2006a): "Belief Attribution and the Falsification of Motive Internalism", in: *Philosophical Psychology* 19 (5): 607-616.
- Cholbi, M. (2006b): "Moral Belief Attribution: A Reply to Roskies", in: *Philosophical Psychology* 19 (5): 629-638.
- Chomsky, N. (1957): *Syntactic Structures*. The Hague/Paris: Mouton.
- Churchland, P. M. (1981): "Eliminative materialism and the propositional attitudes", in: *The Journal of Philosophy* 78: 67-90.
- —— (2005): "Functionalism at Forty", in: *The Journal of Philosophy* 102 (1): 33–50.
- —— (2012): *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. Cambridge, MA.: MIT Press.
- Churchland, P. S., Koch, C. & Sejnowski, T. (1988): "What is computational neuroscience?", in: E. Schwartz, ed., *Computational Neuroscience*. Cambridge, MA.: MIT Press: 46-55.
- Clark, A. (1997): *Being There: Putting Mind, Body, and World Together Again*. Cambridge, MA.: MIT Press.
- —— (2008): *Supersizing the Mind. Embodiment, Action, and Cognitive Extension*. Oxford, New York, NY.: Oxford University Press.
- Clarke, M. (2004): *Reconstructing Reason and Representation*. Cambridge, MA.: MIT Press.

- Cohen, M. & Dennett, D. (2011): "Consciousness cannot be separated from function", in: *Trends in Cognitive Sciences* 15 (8): 358-364.
- Cosmides, L., Tooby, J., & Barkow, J. H. (1992): "Introduction: Evolutionary psychology and conceptual integration", in J. Barkow, L. Cosmides & J. Tooby, eds.: *The adapted mind: Evolutionary psychology and the generation of culture*. New York, NY.: Oxford University Press: 3-15.
- Cottingham, J., Stoothoff, R., Murdoch, D. (1984): *The Philosophical Writings of Descartes*, 2 vols.. Cambridge: Cambridge University Press.
- Cottingham, J., Stoothoff, R., Murdoch, D., Kenny, A. (1991): *The Philosophical Writings of Descartes, Vol. III: The Correspondence*. Cambridge: Cambridge University Press.
- Crane, T. (2001): "Intentional Objects", in: *Ratio* 14: 336-349.
- —— (2013): *The Object of Thought*. Oxford: Oxford University Press.
- Craver, C. (2009): "Mechanisms and Natural Kinds", in: *Philosophical Psychology* 22: 575–594.
- Cummins, R. (1983): *The Nature of Psychological Explanation*. Cambridge, MA.: MIT Press.
- —— (1991): *Meaning and Mental Representation*. Cambridge, MA.: MIT Press.
- —— (2000): ""How does it work" vs "What are the laws?": Two conceptions of psychological explanation", in: F. Keil & R. Wilson, eds., *Explanation and Cognition*. Cambridge, MA.: MIT Press: 117 – 144.
- Cussins, A. (2003): "Content, Conceptual Content and Nonconceptual Content", in: Y. Gunther, ed., *Essays on Nonconceptual Content*. Cambridge, MA.: MIT Press: 133-163.
- D'Esposito, M. & Wills, H. (2000): "Functional Imaging of Neurocognition", in: *Semin. Neurol. 20* (4): 487-498.
- Damasio, A. (1994): *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY.: Grosset/Putnam.
- —— (1996): "The somatic marker hypothesis and the possible functions of the prefrontal cortex", in: *Transactions of the Royal Society (London)* 351: 1413–1420.
- Danto, A. (1973): *Analytical Philosophy of Action*. Cambridge: Cambridge University Press.

- —— (1981): *The Transfiguration of the Commonplace: A Philosophy of Art*. Harvard University Press.
- Danziger, S., Levav, J., Avnaim-Pesso, L. (2011): "Extraneous Factors in Judicial Decisions", in: *Proceedings of the National Academy of Sciences of the United States of America* 108 (17): 6889-6892.
- Davidson, D. (1980): *Essays on Actions and Events*. Oxford: Clarendon Press.
- —— (2001a): *Inquiries into Truth and Interpretation*, 2nd ed.. Oxford: Clarendon Press.
- —— (2001b): *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press.
- —— (2004): *Problems of Rationality*. Oxford: Clarendon Press.
- Davidson, D., McKinsey, J. & Suppes, P. (1955): "Outlines of a Formal Theory of Value", in: *Philosophy of Science* 22: 140-160.
- Davidson, D., Suppes, P. und Siegel, S. (1957): *Decision Making: An experimental approach*. Stanford, CA.: Stanford University Press. Reprinted as a Midway Reprint (1977). Chicago: University of Chicago Press.
- Dayan, P. & Abbott, L. (2001): *Theoretical Neuroscience*. Cambridge MA.: The MIT Press.
- Dennett, D. (1971): "Intentional Systems", in: *Journal of Philosophy* 68 (4): 87–106.
- —— (1987): *The Intentional Stance*. Cambridge, MA.: MIT Press.
- —— (1990): "Quining Qualia", in: W. Lycan, ed., *Mind and Cognition*. Oxford: Blackwell: 519–548.
- —— (1991a): *Consciousness Explained*. Boston: Little, Brown and Company.
- —— (1991b): "Real Patterns", in: *The Journal of Philosophy* 88 (1): 27-51.
- —— (2007): "Philosophy as naïve anthropology: Comment on Bennett and Hacker", in: M. Bennett, D. Dennett, P. Hacker, & J. Searle, eds., *Neuroscience and Philosophy: Brain, Mind, and Language*. New York, NY.: Columbia University Press: 73-95.
- Deonna, J.A. & Teroni, F. (2012): *The Emotions: A Philosophical Introduction*. London and New York: Routledge.
- Deroy, O. (2015): "Multisensory Perception and Cognitive Penetration", in: J. Zeimbekis, A. Raftopoulos, eds., *The Cognitive Penetrability of Perception*. Oxford: Oxford University Press: ch. 5.

- Descartes, R. (1965): *Discourse on Method, Optics, Geometry, and Meteorology*, transl. by P. Olscamp. Indianapolis: Bobbs-Merrill.
- Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G (1992): "Understanding motor events: a neurophysiological study", in: *Experimental Brain Research* 91: 176-180.
- Dretske, F. (1981): *Knowledge and the Flow of Information*. Cambridge, MA.: MIT Press.
- —— (1986): "Misrepresentation", in: R. Bogdan, ed., *Belief*. Oxford: Oxford University Press: 17-36.
- —— (1988): *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA.: MIT Press.
- Duhem, P. (1906/1954): *The Aim and Structure of Physical Theory*. Princeton, NJ.: Princeton University Press.
- Edmonds, D. & Warburton, N., eds. (2015): *Philosophy Bites Again*. Oxford: Oxford University Press.
- Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996): *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA.: MIT Press.
- Ekman, P. (1999): "Basic Emotions", in: T. Dalgleish and M. Power, eds., *Handbook of Cognition and Emotion*. Sussex: John Wiley & Sons Ltd.: 45-60.
- Eskine, K.J, Kacinik, N.A., and Prinz, J.J (2011): "A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment", in: *Psychological Science* 22 (3): 295-299.
- Evans, G. (1982): *The Varieties of Reference*, ed. by J. McDowell. Oxford: Oxford University Press.
- Ewert J.-P. (1970): "Neural mechanisms of prey-catching and avoidance behaviour in the toad (*Bufo bufo* L.)", in: *Brain Behav. Evol.* 3 (1-4): 36-56.
- Ewert J.-P., Schürg-Pfeiffer E., and Schwippert W.W. (1996): "Influence of pretectal lesions on tectal responses to visual stimulation in anurans – field potential, single neuron and behaviour analyses", in: *Acta Biologica Hungarica* 47 (1-4): 89-111.
- Falkenburg, B. (2012): *Mythos Determinismus: Wieviel erklärt uns die Hirnforschung?* Heidelberg: Springer.

- Feyerabend, P. (1962): "Explanation, Reduction and Empiricism", in: H. Feigl and G. Maxwell, eds., *Scientific Explanation, Space, and Time (Minnesota Studies in the Philosophy of Science, Volume III)*. Minneapolis: University of Minneapolis Press: 28–97.
- Field, H. (1975): "Conventionalism and Instrumentalism in Semantics", in: *Nous* 9 (4): 375-405.
- Flanagan O. (1991): *The Science of the Mind*, 2nd Edition. Cambridge, MA.: MIT Press.
- Fodor, J. (1974): "Special Sciences (Or: The Disunity of Science as a Working Hypothesis)", in *Synthese* 28 (2): 97–115.
- —— (1975): *The Language of Thought*. New York, NY.: Thomas Crowell and Cambridge, MA.: Harvard University Press.
- —— (1985): "Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum", in: *Mind* 94 (373): 76-100.
- —— (1989): *Psychosemantics*. Cambridge, MA.: MIT Press.
- —— (1991): "A Modal Argument for Narrow Content", in: *Journal of Philosophy* 88: 5–26.
- —— (1994): *The Elm and the Expert*. Cambridge, MA.: MIT Press.
- —— (1998): *Concepts: Where Cognitive Science Went Wrong*. New York, NY.: Oxford University Press.
- —— (2001): "Language, Thought and Compositionality", in: *Mind and Language* 16: 1–15.
- —— (2008): *LOT2: The Language of Thought Revisited*. Oxford: Oxford University Press.
- Fodor, J. & Pylyshyn, Z. (1988): "Connectionism and Cognitive Architecture: A Critical Analysis", in: S. Pinker and J. Mehler, eds., *Connections and Symbols (A Cognition Special Issue)*. Cambridge, MA.: MIT Press: 3-71.
- Føllesdal, D. (1975): "Meaning and Experience", in: S. Guttenplan, ed., *Mind and Language. Wolfson College Lectures 1974.* Oxford: Oxford University Press: 26–44.
- Frege, G. (1892): "Sinn und Bedeutung", in: *Zeitschrift für Philosophie und philosophische Kritik*, im Verein mit mehreren Gelehrten vormals herausgegeben von. Dr. J. H. Fichte und Dr. H. Ulrici, redigirt von Dr. Richard Falckenberg, Professor der

Philosophie in Erlangen. Neue Folge Hundertster Band. Lepizig: Verlag von C.E.M. Pfeffer.

- Fridland, E. (2015): "Skills, Nonpropositional Thought, and the Cognitive Penetrability of Perception", in: *Journal for General Philosophy of Science* 46 (1): 105-120.
- Friston K. (2010): "The free-energy principle: a unified brain theory?", in: *Nat Rev Neurosci* 11(2):127-138.
- Gallagher, S. (2003): "Phenomenology and Experimental Design: Toward a Phenomenologically Enlightened Experimental Science", in: *Journal of Consciousness Studies* 10 (9-10): 85-99.
- —— (2012): *Phenomenology*. New York, NY.: Palgrave Macmillan.
- Gazzaniga, M., ed. (1995): *The cognitive neurosciences*. Cambridge, MA.: MIT Press.
- Gigerenzer, G. (2008): "Moral Intuition = Fast and frugal Heuristics?", in: W. Sinnott-Armstrong, ed, *Moral psychology, Volume 2: The cognitive science of morality: Intuition and diversity*. Cambridge, MA.: MIT Press: 1-26.
- Gillett, C. (unpublished): "The Truth of the Completeness of Physics is an Open Empirical Question: The Local Battles of Mutualism and Fundamentalism".
- Gillett, C. & Loewer, B., eds. (2001): *Physicalism and Its Discontents*. Cambridge, MA.: Cambridge University Press.
- Gladziejewski, P. (2015): "Action guidance is not enough, representations need correspondence too: A plea for a two-factor theory of representation", in: *New Ideas in Psychology*, doi:10.1016/j.newideapsych.2015.
- Glennan, S. (1996): "Mechanisms and the Nature of Causation", in: *Erkenntnis* 44: 49-71.
- Gold, I. & Stoljar, D. (1999): "A neuron doctrine in the philosophy of neuroscience", in: *Behavioural and Brain Sciences* 22: 809-869.
- Goldman, A. (2012): "Theory of Mind", in: E. Margolis, R. Samuels & S. Stich, eds., *The Oxford Handbook of Philosophy of Cognitive Science*. New York, NY.: Oxford University Press: 402–424.
- Goodman, N. (1972): *Problems and projects*. Indianapolis/New York: Bobbs-Merrill.
- —— (1983): *Fact, Fiction and Forecast,* 4th Edition. Cambridge, MA.: Harvard University Press.

- Gould, S. J., & Lewontin, R. C. (1979): "The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme", in: *Proceedings of the Royal Society* B205: 581–598.
- Greene, J. (2015): "The rise of moral cognition", in: *Cognition* 135: 39-42.
- Grèzes J., Armony J., Rowe J., & Passingham R. (2003): "Activations related to 'mirror' and 'canonical' neurones in the human brain: an fMRI study", in: *NeuroImage* 18: 928-937.
- Grice, P. (1989): *Studies in the Way of Words*. Cambridge, MA.: Harvard University Press.
- Griffin, R. & Baron-Cohen, S. (2002): "The Intentional Stance: Developmental and Neurocognitive Perspectives", in: A. Brook, & D. Ross, eds., *Daniel Dennett. Contemporary Philosophy in Focus*. New York, NY.: Cambridge University Press: 83-116.
- Gunther, Y., ed. (2003): *Essays on Nonconceptual Content*. Cambridge, MA.: MIT Press.
- Hacking, I. (1999): *The Social Construction of What?* Cambridge, MA.: Harvard University Press.
- —— (2007): "Natural Kinds: Rosy Dawn, Scholastic Twilight", in: A. O'hear, ed., *Philosophy Of Science*. Cambridge: Cambridge University Press: 203-239.
- Harris, S., Sheth, S. & Cohen, M. (2008): "Functional Neuroimaging of Belief, Disbelief and Uncertainty", in: *Ann. Neurol.* 63 (2): 141–147.
- Haugeland, J. (1981): "Semantic Engines: An Introduction to Mind Design", in: J. Haugeland, ed., *Mind Design. Philosophy, Psychology, Artificial Intelligence*. Cambridge, MA.: MIT Press: 1–34.
- —— (1985): *Artificial Intelligence: The Very Idea*. Cambridge, MA.: MIT Press.
- —— (1995): "Mind Embodied and Embedded", in: Y. Houng, J. Ho, eds., *Mind and Cognition*. Taipei: Academia Sinica: 207-237.
- —— (2003): "Syntax, Semantics, Physics", in: J. Preston and J. Bishop, eds., *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford: Oxford University Press: 379–392.
- Haynes, J. D. and Rees, G. (2006): "Decoding mental states from brain activity in humans", in: *Nat. Rev. Neurosci. 7*, 523–534.

- Hazlett, A. (2013): *A Luxury of the Understanding: On The Value of True Belief*. Oxford: Oxford University Press.
- Healey, R. (2013): Review of "Scientific Metaphysics", in: *Notre Dame Philosophical Reviews*, online version, http://ndpr.nd.edu/news/41185-scientific-metaphysics (accessed on August 26th 2013).
- Hebb, D.O. (1949): *The Organization of Behavior*. New York: Wiley & Sons.
- Hedden, T., Ketay, S., Aron, A., Markus, H. R. & Gabrieli, J. D. E. (2008): "Cultural influences on neural substrates of attentional control", in: *Psychological Science 19* (1): 12–17.
- Heil, J. (2000): *Philosophy of Mind: A Contemporary Introduction* (Reprint). London/NY: Routledge.
- Henrich, J., Heine, S. J. & Norenzayan, A. (2010): "The weirdest people in the world?", in: *Behavioural and Brain Sciences 33* (2–3): 61–153.
- Hoff, W. D., van der Horst, M. A., Nudel, C. B., & Hellingwerf, K. J. (2009): "Prokaryotic phototaxis", in: *Methods in Molecular Biology* 571: 25–49.
- Holmes, N., Calvert, G. & Spence, C. (2009): "Multimodal Integration", in: M. Binder, N. Hirokawa, & U. Windhorst, eds., *Encyclopedia of Neuroscience*. Berlin: Springer: 2457-2461.
- Horberg, E. J., Oveis, C., Keltner, D., and Cohan, A. B. (2009): "Disgust and the Moralization of Purity", in: *Journal of Personality and Social Psychology* 97 (6): 963-976.
- Horberg, E. J., Oveis, C., and Keltner, D. (2011): "Emotions as Moral Amplifiers: An Appraisal Tendency Approach to the Influences of Distinct Emotions upon Moral Judgment", in: *Emotion Review* 3 (3): 237-244.
- Humeny, C., Kelly, D. & Brook, A. (2012): "Further routes to psychological constructionism", in: *Behavioural and Brain Sciences* 35: 153–154.
- Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J. & Rizzolatti, G. (1999): "Cortical Mechanisms of Human Imitation", in: *Science* 286 (5449): 2526–2528.
- Inbar, Y., Pizarro D., and Bloom, P. (2012): "Disgusting Smells Cause Decreased Liking of Gay Men", in: *Emotion* 12 (1): 1-5.
- Isen, A. & Levin, P. (1972): "Effect of Feeling Good on Helping: Cookies and Kindness", in: *Journal of Personality and Social Psychology* 21: 384-388.

- Jackson, F. (1982): "Epiphenomenal Qualia", in: *Philosophical Quarterly* 32: 127–136.
- Jacobson, A. (2013): *Keeping the World in Mind. Mental Representations and the Sciences of the Mind*. Basingstoke, Hampshire: Palgrave Macmillan.
- Jamieson, D. (2009): "What do animals think?", in: R. Lurz, ed., *The Philosophy of Animal Minds*. Cambridge, New York, NY.: Cambridge University Press: 15-34.
- Jones A. and Fitness, J. (2008): "Moral Hypervigilance: The Influence of Disgust Sensitivity in the Moral Domain", in: *Emotion* 8 (5): 613-627.
- Jones, T. (2004): "Special Sciences: Still a flawed argument after all these years", in: *Cognitive Science* 28: 409-432.
- Kant, I. (2011): *Groundwork of the Metaphysics of Morals: A German-English Edition*, ed. and transl. by M. Gregor & J. Timmermann. Cambridge: Cambridge University Press.
- Kanwisher, N. (2001): "Neural events and perceptual awareness", in: *Cognition* 79: 89–113.
- Kawase, H., Okata, Y. & Ito, K. (2013): "Role of Huge Geometric Circular Structures in the Reproduction of a Marine Pufferfish", in: *Scientific Reports* 3: 2106.
- Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. (2008): "Identifying natural images from human brain activity", in: *Nature* 452: 352–355.
- Kenny, A. (1971): "The homunculus fallacy", in: M. Grene & I. Prigogine, eds., *Interpretations of Life and Mind*. New York, Humanities Press: 155–165.
- —— (1984): "Intentionality: Aquinas and Wittgenstein", in: *The Legacy of Wittgenstein*. Oxford: Basil Blackwell: 61-76.
- Kihlstrom, J. F. (2010): "Social neuroscience: The footprints of Phineas Gage", in: *Social Cognition* 28 (6): 757–82.
- Kim, J. (1993): *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.
- Kincaid, H. & Sullivan, J. (2014): *Classifying Psychopathology: Mental Kinds and Natural Kinds*. Cambridge, MA.: MIT Press.
- Klima, G., ed. (2014): *Intentionality, Cognition and Mental Representation in Medieval Philosophy*. New York, NY.: Fordham University Press.
- Knobe, J. (2015): "Philosophers are doing something different now: Quantitative Data", in: *Cognition* 135: 36-38.

- Kosslyn, S. M. & Andersen, R. A., eds. (1992): *Frontiers in cognitive neuroscience.* MIT Press.
- Kosslyn, S. M. & Koenig, O. (1995): *Wet mind: The new cognitive neuroscience*. The Free Press.
- Kripke, S. (1979): "A puzzle about belief", in A. Margalit, ed., *Meaning and Use*. Dordrecht: Reidel: 239-283.
- —— (1980): *Naming and Necessity*. Oxford: Blackwell.
- —— (1982): *Wittgenstein on Rules and Private Language*. Cambridge, MA.: Harvard University Press.
- Kuhn, T. (1962): *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- —— (2000): *The Road Since Structure*, ed. by J. Conant & J. Haugeland. Chicago: University of Chicago Press.
- Kusch, M. (1999): *Psychological Knowledge: A Social History and Philosophy*. New York, NY.: Routledge.
- Labuschagne, W. & Heidema, J. (2005): "Natural and artificial cognition: On the proper place of reason", in: *South African Journal of Philosophy* 24 (2): 137-149.
- Larmer, R. (1986): "Mind-body interactionism and the conservation of energy", in: *International Philosophical Quarterly* 26: 277-285.
- Leitgeb, H. (2003): "Nonmonotonic reasoning by inhibition nets II", in: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11, Supplement 2: 105-135.
- Levine, J. (1987): "The Nature of Psychological Explanation by Robert Cummins: A Critical Notice", in: *The Philosophical Review* 96 (2): 249-274.
- Levy, N. (2004): "Evolutionary Psychology, Human Universals, and the Standard Social Science Model", in: Biology and Philosophy 19: 459-472.
- Lewis, C. (1929): *Mind and the World Order*. New York, NY.: Charles Scribner's Sons.
- Lewis, D. (1972): "Psychophysical and Theoretical Identifications", in: *Australasian Journal of Philosophy* 50: 249–58.
- —— (1981): "What puzzling Pierre believes", in: *Australasian Journal of Philosophy* 59: 283-289.
- —— (1983a): "Extrinsic Properties", in: *Philosophical Studies* 44: 197–200.

- —— (1983b): *Philosophical Papers Volume 1*. Oxford: Oxford University Press.
- Libet, Benjamin (1985): "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action", in: *The Behavioural and Brain Sciences VIII*, 529-539.
- Libet, B., Gleason, C. A., Wright, E. W. & Pearl, D. K. (1983): "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act", in: *Brain* 106: 623–642.
- Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., & Barrett, L.F. (2012): "The brain basis of emotion: A meta-analytic review", in: *Behavioural and Brain Sciences 35*: 121-143.
- Lowe, E. (2000): "Causal Closure Principles and Emergentism", in: *Philosophy* 75 (294): 571-586.
- Lycan, W. (2000): *Philosophy of Language: A Contemporary Introduction*. London/New York: Routledge.
- MacDonald, P. (2012): *Languages of Intentionality: A Dialogue Between Two Traditions of Consciousness*. London/New York: Continuum.
- Marr, D. & Poggio, T. (1977): "From Understanding Computation to Understanding Neural Circuitry", in: *Neurosciences Res. Prog. Bull.* 15 (3): 470-488. Also available as M.I.T. AIM-357.
- McCloud, Scott (1993): *Understanding Comics: The Invisible Art*. New York, NY.: HarperCollins.
- McCulloch, W. & Pitts, W. (1943): "A logical calculus immanent in nervous activity", in: *Bulletin of Mathematical Biophysics* 5: 115-133.
- McDowell, J. (1994): *Mind and World.* Cambridge, MA.: Harvard University Press.
- Menary, R., ed. (2010): *The Extended Mind*. Cambridge, MA./London: MIT Press.
- Mendonça, W. (2010): "Mental Causation and the Causal Completeness of Physics", in: *Principia* 6 (1): 121-132.
- Mercier, H. & Sperber, D. (2011): "Why do humans reason? Arguments for an argumentative theory", in: *Behavioural and Brain Sciences* 34 (2): 57–111.
- Millikan, R. (1989): "Biosemantics", in *Journal of Philosophy* 86: 281–97.
- —— (1993): *White Queen Psychology and Other Essays for Alice*. Cambridge, MA.: MIT Press.

- Mineka, S. & Cook, M. (1989): "Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus-monkeys", in: *Journal of Abnormal Psychology* 98 (4): 448-459.
- Montero, B. (2006): "What does the conservation of energy have to do with physicalism?", in: *Dialectica* 60 (4): 383-396.
- Morris, C. (1938): "Foundations of the theory of signs," in O. Neurath, R. Carnap and C. Morris, eds., *International Encyclopaedia of Unified Science I*. Chicago: University of Chicago Press: 77–138. Reprinted in C. Morris (1971): *Writings on the general theory of signs*. The Hague: Mouton.
- Morrison, M. (2000): *Unifying Scientific Theories: Physical Concepts and Mathematical Structures*. Cambridge: Cambridge University Press.
- Nachev, P. & Hacker, P. (2014): "The neural antecedents to voluntary action: A conceptual analysis", in: *Cogn. Neurosci.* 5 (3-4): 193-208.
- Nagel, T. (1961): *The Structure of Science*. London: Routledge and Kegan Paul.
- —— (1974): "What Is It Like to Be a Bat?", in: *The Philosophical Review* 83 (4): 435–450.
- —— (2012): *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*. New York, NY.: Oxford University Press.
- Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009): "Bayesian reconstruction of natural images from human brain activity", in: *Neuron* 63: 902–915.
- Nathan, M. & Del Pinal, G. (2015): "Mapping the mind: Bridge laws and the psycho-neural interface", in: *Synthese* May 2015, DOI 10.1007/s11229-015-0769-2.
- Neander, K. (1995): "Misrepresenting & Malfunctioning", in: *Philosophical Studies* 79: 109–41.
- Neumann, John von & Morgenstern, Oskar (1944): *Theory of Games and Economic Behavior*. Princeton, NJ.: Princeton University Press.
- Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J. (2011): "Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies", in: *Current Biology 21*: 1641-1646.
- Noë, A. (2004): *Action in Perception*. Cambridge, MA/London: MIT Press.
- Okano, H., Hirano T. & Balaban, E. (2000): "Learning and Memory", in: *Proceedings of the National Academy of Sciences of the United States of America* 97 (23): 12403-12404.

- Oppenheim, P. & Putnam, H. (1958): "The unity of science as a working hypothesis", in H. Feigl et al., eds., *Minnesota Studies in the Philosophy of Science*, Vol. 2. Minneapolis: Minnesota University Press.
- Overgaard, M., ed. (2015): *Behavioural Methods in Consciousness Research*. Oxford: Oxford University Press.
- Papineau, D. (1991): "The Reason Why: Response to Crane", in: *Analysis* 51 (January): 37-40.
- Peirce, A. (1998): *The Essential Peirce,* Volume 2. Eds. Peirce Edition Project. Bloomington, IN.: Indiana University Press.
- Penfield, W. & Rasmussen T. (1950): *The Cerebral Cortex of Man. A Clinical Study of localisation of Function*. New York, NY.: Macmillan.
- Pinker, S. (1997): *How the Mind Works*. New York, NY.: W. W. Norton & Company.
- Pitkin, H. (1967): *The Concept of Representation*. Berkeley: University of California Press.
- Place, U. T. (2002): "Is Consciousness a Brain Process?", in: D. Chalmers, ed., *Philosophy of Mind – Classical and Contemporary Readings*. Oxford: Oxford University Press: 55–60.
- Popper, K. (1959): *The Logic of Scientific Discovery*. London: Hutchinson.
- Prinz, J. (2004): *Gut Reactions: A Perceptual Theory of Emotion*. Oxford: Oxford University Press.
- Putnam, H. (1975): "The Meaning of 'Meaning'," in: ibid., *Mind, Language and Reality (Philosophical Papers, Volume 2)*. Cambridge: Cambridge University Press: 215-271.
- —— (1981): *Reason, Truth, and History*. Cambridge: Cambridge University Press.
- Qin, P., Duncan, N. & Northoff, G. (2013): "How is our self related to the brain midline regions?", in: *Frontiers in Human Neuroscience* 7: 909.
- Quine, W.V.O. (1956): "Quantifiers and propositional attitudes", in: *Journal of Philosophy* 53: 177-187.
- —— (1960): *Word and Object*. Cambridge, MA.: MIT Press.
- —— (1969): *Ontological Relativity and Other Essays*. New York, NY.: Columbia University Press.
- —— (1980): *From a Logical Point of View.* Cambridge, MA.: Harvard University Press.

- —— (1993): "In Praise of Observation Sentences", in: *Journal of Philosophy* 60 (3): 107-116.
- —— (2008): *Confessions of a Confirmed Extensionalist and Other Essays*, ed. by D. Føllesdal and D. Quine. Harvard: Harvard University Press.
- Quine, W.V.O. & Ullian, J. (1970): *The Web of Belief*. New York, NY.: Random House.
- Radman, Z., ed. (2012): *Knowing without thinking: mind, action, cognition and the phenomenon of the background*. Basingstoke, Hampshire/New York, NY.: Palgrave Macmillan.
- Ramsey, F. (1931): *The Foundations of Mathematics and Other Logical Essays*, ed. by R. Braithwaite. London: Routledge & Kegan, Paul, Trench, Trubner & Co./New York, NY.: Harcourt, Brace and Company.
- Ramsey, W. (2007): *Representation reconsidered*. Cambridge: Cambridge University Press.
- Rauss, K., Schwartz, S. & Pourtois G. (2011): "Top-down effects on early visual processing in humans: a predictive coding framework", in: *Neurosci. Biobehav. Rev.* 35 (5): 1237-53.
- Rechenauer, M. (1994): *Intentionaler Realismus und Externalismus: Beiträge zur Diskussion des Individualismus in der analytischen Philosophie des Geistes*. Würzburg: Königshausen und Neumann.
- —— (1997): "Individualism, individuation and that-clauses", in: *Erkenntnis* 46 (1): 49-67.
- Reig, R. and Silberberg, G. (2014): "Multisensory Integration in the Mouse Striatum", in: *Neuron* 83 (5): 1200-12.
- Roskies, A. (2003): "Are ethical judgments intrinsically motivational? Lessons from 'acquired sociopathy'[1]", in: *Philosophical Psychology* 16 (1): 51-66.
- —— (2006): "Patients With Ventromedial Frontal Damage Have Moral Beliefs", in: *Philosophical Psychology* 19 (5): 617-627.
- Russell, B. (1918): "The Philosophy of Logical Atomism", in: *The Monist* 1918. Reprinted in ibid. (1956): *Logic and Knowledge: Essays 1901–1950*, ed. by R. Marsh, London: Unwin Hyman: 177–281 and in D. Pears, ed. (1985): *The Philosophy of Logical Atomism*, La Salle, IL: Open Court: 35–155.
- —— (1950): *Unpopular Essays*. London/New York, NY.: Routledge.

- —— (1973) [1935]: *In Praise of Idleness and other essays*. London: Unwin Books.
- Ryle, G. (1949): *The Concept of Mind*. London: Hutchinson.
- Sauer, H. (2012): "Educated Intuitions. Automaticity and Rationality in Moral Judgment", in: *Philosophical Explorations* 15 (3): 255-275.
- Saussure, F. de (1983): *Course in General Linguistics*, transl. by R. Harris. London: Duckworth.
- Schleim, S. (2011): *Die Neurogesellschaft – Wie die Hirnforschung Recht und Moral herausfordert*. Hannover: Heise Verlag.
- Schnall, S., Benton, J., Harvey, S. (2008a): "With a Clean Conscience. Cleanliness Reduces the Severity of Moral Judgments", in: *Psychological Science* 19 (12): 1219-1222.
- Schnall, S., Haidt, J., and Jordan, A.H. (2008b): "Disgust as Embodied Moral Judgment", in: *Personality and Social Psychology Bulletin* 34 (8): 1096-1109.
- Schrödinger, E. (1992): *What Is Life? with Mind and Matter and Autobiographical Sketches*. Cambridge University Press.
- Searle, J. (1979): "What is an intentional state?", in: *Mind* 88: 74-92.
- —— (1980): "Minds, brains and programs", in: *The Behavioural and Brain Sciences* 3 (3): 417-24.
- —— (1983): *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.
- —— (1989): "Artificial Intelligence and the Chinese Room: An Exchange", in: *New York Review of Books* 36 (2): 44-45.
- —— (1990): "Is the Brain a Digital Computer?", in: *Proceedings and Addresses of the American Philosophical Association* 64 (3): 21-37.
- —— (1992): *The Rediscovery of the Mind*. Cambridge, MA.: MIT Press.
- —— (1995): *The Construction of Social Reality*. London: Penguin.
- —— (2000): *Mind, Language and Society: Philosophy in the Real World*. London: Phoenix.
- —— (2010): "Why Dualism (and Materialism) Fail to Account for Consciousness", in R. Lee, ed., *Questioning Nineteenth Century Assumptions about Knowledge, III: Dualism*. New York, NY.: SUNY Press: Section I.
- Selemon, L. (2013): "A role for synaptic plasticity in the adolescent development of executive function", in: *Translational Psychiatry* 3: e238.

- Sellars, W. (1957): "Intentionality and the Mental – A symposium by correspondence with Roderick Chisholm", in: H. Feigl, M. Scriven & G. Maxwell, eds., *Minnesota Studies in the Philosophy of Science*, Vol. II. Minneapolis: University of Minnesota Press: 507–39.
- —— (1981): "Mental Events", in: *Philosophical Studies* 81: 325–45.
- —— (1997): *Empiricism and the Philosophy of Mind*, with an introduction by Richard Rorty and a study guide by Robert Brandom. Cambridge, MA./London: Harvard University Press.
- Seok, B. (2006): "Diversity and Unity of Modularity", in: *Cognitive Science* 30: 347-380.
- Sera, M., Elieff, C., Forbes, J., Burch, M., Rodríguez, W., Dubois, D. (2002): "When language affects cognition and when it does not: An analysis of grammatical gender and classification", in: *Journal of Experimental Psychology: General* 131 (3): 377-397.
- Shannon, C. (1948): "A Mathematical Theory of Communication", in: *Bell Systems Technical Journal* 27: 279-423, 623-656.
- Shapiro, L. (2010): *Embodied Cognition*. London/New York, NY.: Routledge.
- Shih, J. & Cohen L. (2004): "Cortical reorganization in the human brain: how the old dog learns depends on the trick", in: *Neurology* 63 (10): 1772-3.
- Shoemaker, S. (1996): *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.
- Shope, R. (1999): *The Nature of Meaningfulness: Representing, Powers, and Meaning*. Oxford: Rowman & Littlefield.
- Sie, M. & Wouters, A. (2010): "The BCN Challenge to Compatibilist Free Will and Personal Responsibility", in: *Neuroethics* 3: 121-133.
- Simons, D. & Chabris, C. (1999): "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events", in: *Perception 28*: 1059–1074.
- Skinner, B. F. (1974): *About Behaviorism.* New York, NY.: Vintage.
- Sklar, L. (1967): "Types of inter-theoretic reduction", in: *The British Journal for the Philosophy of Science* 18: 109–124.
- Smith, D. L. (1999): *Freud's Philosophy of the Unconscious (Studies in Cognitive Systems, Vol. 23)*. Dordrecht & Boston: Kluwer Academic Publishers.

- Smolensky, P. (1988): "On the proper treatment of connectionism", in: *Behavioral and Brain Sciences* 11:1-23.
- Smolensky, P., Legendre, G. & Miyata, Y. (1992): "Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition", in: *Tech Report 92-08*, Institute of Cognitive Science, University of Colorado.
- Snyder, L. J. (2006): *Reforming Philosophy: A Victorian Debate on Science and Society*. Chicago: University of Chicago Press.
- Sosa, E. (1970): "Propositional attitudes *de dicto* and *de re*", in: *Journal of Philosophy* 67: 883-896.
- Stapp, H. (2009): "Physicalism versus Quantum Mechanics", in: *Mind, Matter and Quantum Mechanics*. The Frontiers Collection: 245-260.
- Strohminger, N. (2015): "Need for Empirical Recognition", in: *Emotion Review* 05/2015, DOI: 10.1177/1754073915583917.
- Sullivan, J. (2009): "The Multiplicity of Experimental Protocols: A Challenge to Reductionist and Non-Reductionist Models of the Unity of Neuroscience", in: *Synthese* 167: 511-539.
- —— (2010): "A Role for Representation in Cognitive Neurobiology", in: *Philosophy of Science* 77 (5): 875-887.
- —— (2014): "Is the Next Frontier in Neuroscience a 'Decade of the Mind'?", in: C. Wolfe, ed., *Brain Theory. Essays in Critical Neurophilosophy*. Basingstoke: Palgrave Macmillan: 45-67.
- —— (2015a): "Experimentation in Cognitive Neuroscience and Cognitive Neurobiology", in: J. Clausen & N. Levy, eds., *Springer Handbook of Neuroethics*. Springer: 31-47.
- —— (2015b): "Neuroscientific kinds through the lens of scientific practice", in: C. Kendig, ed., *Natural Kinds and Classification in Scientific Practice*. New York: Routledge: 47-56.
- —— (forthcoming): "Construct Stabilization and the Unity of the Mind-Brain Sciences", in: *Philosophy of Science* 83: 662-673.
- Sunstein, C. (2005): "Moral Heuristics", in: *Behavioral and Brain Sciences* 28 (4): 531-573.
- Tarski, A. (1986): *Collected Papers Vol. 2. 1935-1944*. Basel: Birkhäuser.

- Thagard, P. (2005): *Mind. Introduction to Cognitive Science*, 2nd Ed.. Cambridge, MA.: MIT Press.
- Thelen, E. & Smith, L. (1994): *A dynamic systems approach to the development of cognition and action*. Cambridge, MA.: MIT Press.
- Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., and Dehaene, S. (2006): "Inverse retinotopy: Inferring the visual content of images from brain activation patterns", in: *Neuroimage* 33: 1104–1116.
- Thomasson, A. (1999): *Fiction and Metaphysics*. Cambridge: Cambridge University Press.
- Thompson, M. (2008): *Life and Action. Elementary Structures of Practice and Practical Thought.* Cambridge, MA.: Harvard University Press.
- Tiehen, J. (2015): "Explaining Causal Closure", in: *Philosophical Studies* 172 (9): 2405-2425.
- Tretter, F., Winterer, G., Gebicke-Haerter, P., Mendoza, E., eds. (2010): *Systems Biology in Psychiatric Research. From High-Throughput Data to Mathematical Modelling*. Weinheim: Wiley.
- Triskiel, J. (2016): "Psychology instead of Ethics? Why psychological research is important but cannot replace ethics", in: C. Brand, ed., *Dual Process Theories in Moral Psychology*. Wiesbaden: Springer: 77-98.
- Tversky, A. & Kahneman, D. (1974): "Judgment under Uncertainty: Heuristics and Biases", in: *Science* 185 (4157): 1124-1131.
- Twardowski, K. (1977): *On the Content and Object of Presentations*, transl. by R. Grossmann. The Hague: M. Nijhoff.
- Valdesolo P. & de Steno, D. (2006): "Manipulations of emotional context shape moral judgment", in: *Psychological Science* 17 (6): 476-477.
- Valenstein, E. (1973): *Brain Control*. New York, NY.: Wiley.
- Vasilyev, V. (2009): "The Hard Problem of Consciousness and Two Arguments for Interactionism", in: *Faith and Philosophy* 26 (5): 514-526.
- Wachowitz S. & Ewert J.-P. (1996): "A key by which the toad's visual system gets access to the domain of prey", in: *Physiol. Behav.* 60 (3): 877-887.
- Wachter, D. von (2006): "Why the Argument from Causal Closure against the Existence of Immaterial Things is Bad", in: H. Koskinen, R. Vilkko & S. Philström, eds., *Science - A Challenge to Philosophy?* Frankfurt/M.: Peter Lang: 113-124.

- Waskan, J. (2006): *Models and Cognition: Prediction and Explanation in Everyday Life and in Science*. Cambridge, MA.: MIT Press.
- Wheatly, T. & Haidt, J. (2005): "Hypnotic Disgust makes Moral Judgments more severe", in: *Psychological Science* 16 (10): 780-784.
- Williams, M. (1999): *Wittgenstein, Mind and Meaning. Toward a social conception of mind*. New York, NY.: Routledge.
- Wilson, R. & Keil, F., eds. (1999): *MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA.: MIT Press.
- Wittgenstein, L. (1953): *Philosophical Investigations*, ed. by G.E.M. Anscombe and R. Rhees, transl. by G.E.M. Anscombe. Oxford: Blackwell.
- —— (1961): *Tractatus Logico-Philosophicus*, transl. by D.F. Pears and B.F. McGuinness. New York, NY.: Humanities Press.
- Woodward, J. (2003): *Making Things Happen*. Oxford University Press.
- Yates, D. (2009): "Emergence, Downward Causation and the Completeness of Physics", in: *The Philosophical Quarterly* 59 (234): 110-131.
- Zeglen, U., ed. (1991): *Donald Davidson: Truth, Meaning and Knowledge*. London: Routledge.
- Zehetleitner, M., & Schönbrodt, F. (2013): "When misrepresentation is successful", in: T. Breyer, ed., *Epistemological dimensions of evolutionary psychology*. New York, NY.: Springer: 197-222.
- Zehetleitner, M. (forthcoming): [tba].

Copyright of figures

Figure 2 (p. 26): Permission obtained from *HarperCollins Publishers*.