A Therapeutic Elimination of "Belief" and "Desire" from Causal Accounts of Action

Mark Curtis

University of East Anglia

School of Politics, Philosophy, Language and Communication Studies

Department of Philosophy

This thesis is submitted for the degree of Doctor of Philosophy

June 2016

Acknowledgements and Dedication

My gratitude, firstly, to my thesis supervisors Eugen Fischer and John Collins. Eugen, as my unstinting primary supervisor, gave short shrift to my occasionally woolly thinking, detached me from my own dogmas and challenged me to raise the standard of my scholarship. Any academic merit that this thesis displays I owe to his guidance. Thanks also to colleagues, faculty and staff at the University of East Anglia, for revealing conversations in both formal and informal settings. Among others I must mention Lewis Clarke, Jessica Dollman, Gareth Jones, Janis Loschmann, Andrei Nasta, Liz McKinnell, Mihai Ometita, Sylvia Panizza, Fiona Roxburgh, Maria Serban, Sidra Shahid, Simon Summers and Odai al Zoubi. I am incredibly grateful to my wife, Karen, without whose emotional and material support I could never have embarked on this project.

This work is dedicated to the memory of my late father, John Curtis, who imbued in me the sense that life should be marked by continual learning, curiosity and enquiry.

Table of Contents

Introduction

Abstract: This introduction sets out the objectives, topic, method and structure of this thesis. I describe ***philosophical folk psychology*** *and the roles that it is presumed to play in* ***action choice****,* ***interpersonal understanding*** *and* ***reason giving****. Philosophical folk psychology – particularly when expressed as* ***belief-desire psychology*** *– is suggested by some as a way to describe all three of these phenomena under a single model. I argue, however, that this comes at the cost of a number of unwarranted commitments which give rise to* ***philosophical problems****. I introduce a handful of influential thinkers who have advanced folk psychological positions and also some contemporary examples of philosophers addressing problems arising directly from it. I then introduce the* ***diagnostic-therapeutic*** *intent of this thesis, grounded in a reading of Wittgenstein's approach to philosophy through the later work of Gordon Baker. Thereafter I set out the two-part structure of the thesis and briefly outline the chapters.*

The human being is the best picture of the human soul.

Wittgenstein, *Culture and Value*
MS 131 80[1]

0.1 Research Questions and Objectives

For millennia, western philosophy engaged with questions of **action**, and **action choice**, **interpersonal understanding** – the ability that most people, most of the time, have to explain and predict other people's actions – and **reason-giving** – how individuals account for their own actions and action-choices. In pursuit of answers to these, a single idea has become prevalent since the so-called **cognitive revolution** in the philosophy of mind and in psychology (Gardner 1987; Miller 2003) that marked the abandonment of **behaviourism** (particularly in the United States) at the beginning of the 1960s. This is the idea that actions can be correctly explained in terms of the actor's possession of specific *causal* **propositional attitudes**, especially **beliefs and desires**. This so-called **common-sense** or

[1] Cf. *Philosophical Investigations* II, iv, 25: "The human body is the best picture of the human soul."

folk psychology[2] is presumed to reflect the way that non-philosophers and non-psychologists understand and explain action.

In this thesis, I am seeking to *challenge the warrant that the proponents of philosophical folk psychology have for their commitments*. I will justify this challenge by:

1. Elucidating the features of psychological explanation in contemporary cognitive and social psychology and
2. Examining several ways in which beliefs and desires are used in everyday, non-philosophical settings.

Among the questions I pose in the present thesis is *whether or not the assumptions and commitments of classic philosophical folk psychology, of the kind described in section 0.2 below, are warranted*.

I contend that if philosophers pay attention to alternative ways of viewing the three phenomena (action and action-choice, interpersonal understanding and reason-giving), then many **philosophical problems** associated with philosophical folk psychology (section 0.3) need not arise. Support for this contention arises from answers to these **guiding questions**:

GQ1. What roles do attributions of specific beliefs and desires play in the way that contemporary **cognitive psychology** investigates *action choice*? And what role do they play in the models of *interpersonal understanding* developed by contemporary **social psychology**?

GQ2. What roles do attributions of specific beliefs and desires play in everyday discourse? Particularly, are they principally used as part of **causal explanations** and, if not, what other purposes do they serve?

Underpinning the first guiding question is the contention that scientific investigations are concerned with **causal accounts** of their target phenomena (established in Chapter 1). If "belief" and "desire" pick out the causes of action as suggested by philosophical folk psychology, we would expect them to feature in the causal-explanatory models developed by cognitive psychologists. Likewise, if philosophical folk psychology accurately described the basis of interpersonal understanding we would expect these terms, their cognates, or the

[2] "Folk psychology" was originally a pejorative term used by opponents of the idea that the folk view was essentially correct, but has been appropriated by its proponents. Accordingly, I use it here without negative connotations.

concepts underlying "belief" and "desire" to feature prominently in the models of interpersonal understanding developed by social psychology.

To date, very few (if any) philosophers have challenged philosophical folk psychology through an elucidation of how psychologists treat the phenomena. Where such considerations have appeared, philosophical effort has concentrated on a reconciliation of scientific models with the default view in philosophy – see Haselager (1997) and the discussion of the **interface problem** in section 0.3 of this introduction. Philosophers struggle with this and other conceptual issues arising from "belief" and "desire" as terms that *essentially pick out causes of action*. In the first part of this thesis I discuss how the causes of action and our understanding of those causes are explained by science.

The second guiding question has also been neglected. Another assumption of philosophical folk psychology is that the "folk"[3] understand propositional attitude terms such as "belief" and "desire" as picking out **mental states** that stand in a causal relationship to action and action choice. The second part of this thesis addresses whether this is necessarily the way that such terms are used and understood. What other purposes might claims of specific beliefs and desires or descriptions of others involving the terms "belief" and "desire" fulfil? How closely do everyday uses of "belief" and "desire" map onto the similarly named concepts of philosophical folk psychology?

The objectives of this thesis are:

1) To clarify whether the terms "belief" and "desire" pick out the **causes** of action in a way that is central to philosophical folk psychology.
2) To examine how interpersonal understanding is investigated by scientists and how this contrasts with philosophical folk psychology.
3) To ask whether reason-giving should be regarded and as implying the same commitments to the causal efficacy of beliefs and desires and to their central role in interpersonal understanding assumed by philosophical folk psychology.
4) To suggest that if this clarification and the offered alternatives are embraced, then many of the philosophical problems arising from philosophical folk psychology can be avoided.

[3] Presumably meaning everyday language-users freed from philosophical constraints.

0.2 Classic Philosophical Folk Psychology

Philosophical folk psychology is built on a number of assumptions, including:

A1) Individuals are caused to choose and to perform particular actions by the *propositional attitudes*, *especially specific beliefs and desires*, that they hold.

A2) People not engaged in philosophical reflection or psychological investigations (the "folk") understand, explain and predict their own and other people's actions with reference to the propositional attitudes – again, especially beliefs and desires – that they either infer to be held by the actor, or find themselves to hold by introspection.

A3) People give reasons for their own actions and choices based on the introspection of their own propositional attitudes, especially beliefs and desires (reason giving).

A4) These capacities are underpinned by knowledge of a *causal relationship* between beliefs and desires and actions or action choice.

In this list, A1 deals expressly with **action and action-choice**, A2 with **interpersonal understanding**[4] and A3 with **reason giving**. Thanks to the causal relationship implied in the fourth assumption, belief-desire psychology offers a unified account of all three capacities. The causal relationship is enshrined in the **belief-desire law** (Bermúdez 2009: 43) that underpins philosophical folk psychology. Horgan and Woodward (1991: 149) offer a heavily qualified version:

> If someone desires that p, and this desire is not overridden by other desires, and he believes that an action of kind K will bring it about that p, and he believes that such an action is within his power, and he does not believe that some other kind of action is within his power and is a preferable way to bring about that p, then ceteris paribus, the desire and the beliefs will cause him to perform an action of kind K.

Although this formulation states the central role of beliefs and desires and the causal relationship in which these are presumed to stand to behaviour, it is perhaps unsurprising that it is usually abbreviated to something like:

[4] I use this "theory-neutral" term for this social ability in preference to "theory of mind" or "mind-reading".

BELIEF-DESIRE LAW:

> An agent who *desires* to bring about a particular state of affairs (P) and who *believes* that performing a specific action (**Φ**) will bring about P will, *ceteris paribus*, tend to **Φ**.

This form is the belief-desire law for the purposes of the present thesis. The *ceteris paribus* clause is meant to capture all of the vagaries spelled out in the version from Horgan and Woodward (1991). If one wanted to be uncharitable one might be tempted to parody their description of the law as "people will do whatever they believe will fulfil their desires – except when they don't". However, Horgan and Woodward make explicit that the kinds of variables constituting their exceptions are *likely to be further beliefs and desires that take precedence*. The causal roles of beliefs and desires are preserved, even when people act in ways that are at odds with – or in outright contradiction of – their declared beliefs and desires.

Philosophers employ belief-desire psychology, based on the four assumptions and the belief desire law, to formulate answers to a number of questions related to action:

1. How do we explain the individual's choice of particular actions in preference to the innumerable courses of action that might be available in many circumstances?
2. How does interpersonal understanding work? What strategies, abilities, perceptions, intuitions, etcetera are engaged when most people[5], most of the time find other's action choices explicable and, to a large degree, predictable?
3. How are the answers to both of these questions related to the way that individuals account for their own actions and action choices?
4. In what way are the answers to questions A-C related?

Each of these questions overlaps with the empirical investigations of psychology.

If "folk psychology", signified nothing more than the everyday recognition of human and non-human agents as **creatures with minds** who are able to acquire and process information about their situation and for whom information-processing operations lead to the selection of actions, as suggested by Fletcher (1995), it would be hard to question. That most of us appreciate **agents** (our peers and many non-human animals) as significantly

[5] One exception being those people on the more socially disadvantaged reaches of the autistic spectrum.

different from other medium-sized objects in our environment seems uncontroversial[6]. In fact, if anything, we are prone to over-attribute agency, as shown through countless replications and refinements of the effect first investigated by Heider and Simmel (1944) and recently reviewed by Keil and Newman (2015) (See Chapters 3 and 4 for more on this phenomenon).

Philosophical folk psychology, however, goes further. It assigns causal force to the presumed referents of the terms "belief" and "desire". It defines information acquisition and processing as synonymous with "belief" and "desires" as picking out motivating states. It accepts the belief-desire law as the process underlying action choice and "the folk" are assumed to have known this all along[7]. I contend that *these assumptions and commitments are held without warrant* other than their status as the orthodox (default) account of action, interpersonal understanding and reason giving.

In their defence of philosophical folk psychology against potential challenge from **cognitive science** and **cognitive psychology** Horgan and Woodward (1991: 149) define it as:

> A network of principles which constitutes a sort of common-sense theory about how to explain human behaviour. These principles provide a central role to certain propositional attitudes, particularly beliefs and desires.

According to Ratcliffe and Hutto (2007: 2) the "received wisdom" of philosophical folk psychology:

> Encapsulates two key assumptions: (1) that making sense of actions requires interpreting them in terms of reasons composed of various propositional attitudes (at a bare minimum – beliefs and desires) and (2) that this activity is primarily concerned with providing predictions and explanations of actions.

To these two, Ratcliffe and Hutto add a third assumption, namely that such a folk psychology is taken to be the "central, core ability that underlies all interpersonal understanding and interaction, rather than just one among many ingredients of social ability" (ibid.). From this the philosopher deduces that the explanatory success of the

[6] Once again, this ability seems to be inhibited in some autistic subjects.

[7] As we shall see, there is some dispute as to how this knowledge is acquired.

strategy must depend on its being *true* (Fodor 1993) and that *because it is true we are justified in using it to construct philosophically robust accounts of action*.

Stich and Nichols (2003) suggest that the question "what is folk psychology" engenders two possible answers. The first is that folk-psychology is little more than the collection of everyday platitudes about mental states and actions (and their relations) that we all "take for granted". This is encapsulated in this quotation from **David Lewis**:

> Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. Perhaps we can think of them as having the form:
>
> *When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and so probability to be caused thereby to go into so-and-so mental states and produce so-and-so motor responses.*
>
> … Include only platitudes that are common knowledge among us – everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that *names of mental states derive their meaning from these platitudes*.
>
> (Lewis 1980: 212)

This description, although somewhat vague ("so-and-so mental states*"), leads directly to the belief-desire law*. As Stich and Nichols describe it, "…we might think of folk psychology as a set of generalizations that systematizes the platitudes in a perspicuous way. A systematization of that sort might also make it more natural to describe folk psychology as a theory" (Stich and Nichols 2003: 240).

The second possible answer to the question "what is folk psychology" suggested by Stich and Nichols focuses on *interpersonal understanding and on the way that the assumptions and the belief-desire law facilitate the everyday prediction of action*. Stich and Nichols suggest that with circumstantial information individuals are rather good not only at predicting and explaining behaviour but also at attributing mental states to other people. If you asked person A "what does person B *believe* about this situation" and then asked person B "what do you believe about this situation" then, so long as person A was able to observe

what person A could *perceive* about the situation, or had other knowledge of Person B's *information* about the situation, then Person B's *account* of their beliefs is likely to coincide with Person A's *ascription of beliefs* to Person B.

For Churchland (1988: 97) too, philosophical folk psychology is a development of a number of observations about human facility with interpersonal understanding. He writes:

> Consider the considerable capacity that normal humans have for explaining and predicting the behaviour of their fellow humans. We can even explain and predict the psychological states of other humans. We explain their behaviour in terms of their beliefs and desires, and their pains, hopes and fears. We explain their sadness in terms of their disappointment, their intentions in terms of their desires, and their beliefs in terms of their perceptions and inferences.

For this to be possible, Churchland contends, each of us must "…be in command of a rather substantial set of laws or generalizations connecting the various mental states with (1) other mental states, with (2) external circumstances and with (3) overt behaviours," (Ibid.). This entails that such "theoretic" capability is facilitated by "hundreds" of common-sense generalisations such as "persons tend to feel pain at points of recent bodily damage", "persons denied fluids for some time tend to feel thirst" and (a restatement of the belief-desire law) "Persons who want that P, and believe that Q would be sufficient to bring about P, and have no conflicting wants or preferred strategies, will try to bring it about that Q," (Ibid: 98).

Disagreement as to whether the skills inherent in folk-psychological prediction and explanation of action constitute a widely held **theory** (or "**folk theory**") of action, as described by Churchland or are achieved by other, non-theoretic means has given rise to the divergence of two schools of thought in folk psychology (Stich and Nichols 2008). On the one hand, *folk-psychological aptitude is said to rest on the possession and empirical development of a theory, analogous to a scientific theory* and its refinement (Gopnik and Meltzoff 1997; Waismeyer et al. 2014) – so called **"theory-theory"**. Others, frustrated by doubts about the provenance of an innate theory and whether its development can be likened to scientific investigation (an additional putative innate capacity) suggest that *folk psychological proficiency rests on an ability to simulate the mental processes of others* (Goldman 2006; Gordon 2008). This **simulation theory** suggests that we are able to run an "off-line simulation" of other people's beliefs and desires and so predict what they do by

hypothesising how we might act in a similar situation. We only need to imagine ourselves with similar motivations (desires) and information (beliefs). Likewise, the post-hoc explanation of another person's behaviour is a matter of describing what *we* would have needed to believe and desire *in order for it to be rational that we would act in the way that we have witnessed*.

There are a number of variations on simulation theory (described by Gordon, 2008) and many hybrid positions that draw from theory-theory and from simulation-based accounts in varying measures (Stich and Nichols 2008: 390). All depend upon the four assumptions and on the belief-desire law as a *correct account of the causal relation between beliefs and desires* and *action*. The dispute is over how the facility to ascribe beliefs and desires is acquired and developed[8] and not over whether such ascriptions form the basis of interpersonal understanding or whether the ascribed states are the genuine causes of action. People with interpersonal skills are assumed to have these assumptions and the belief-desire law encoded in their brains as a matter of course. These are *information-rich accounts of a near-universal human capability* (Stich and Nichols 2003: 241).

However, if the assumptions are not true, and the belief-desire law not binding, then discussion of whether theory or simulation grounds folk-psychology is moot.

The platitude-based account of folk psychology or of the belief-desire law is *intuitively plausible*. Lewis, once again states their significance:

> The concepts of belief, desire, and meaning are common property. The theory that implicitly defines them had better be common property too. It must amount to nothing more than a mass of platitudes of common sense, though these may be reorganized in perspicuous and unfamiliar ways.
>
> (Lewis 1974: 335)

Lewis (2006), lays out the causal relationship between the elements of philosophical folk psychology and the performance of actions in equally bold terms:

> Folk psychology concerns the causal relations of mental states, perceptual stimuli and behavioural responses. It says how mental states, singly or in

[8] In Chapter 4, section 4.7 of the present thesis I introduce a third approach: Hutto's *Narrative Practice Hypothesis* (Hutto 2008b).

> combination*, are apt for causing behaviour*; and it says how mental states are apt to change under the impact of perceptual stimuli and other mental states. Thus *it associates with each mental state a typical causal role*. … Whenever *M* is a folk-psychological name for a mental state, folk psychology will say that *M* typically occupies a certain causal role: call this the *M*-role. Then we analyse *M* as meaning 'the state that typically occupies the *M*-role. *Folk psychology implicitly defines the term M, and we have only to make that definition explicit*.
>
> (Lewis 2006: 56, emphasis added,)

The final sentence of this quotation epitomises the *essential commitment to the causal meaning* of "belief" and "desire" that marks philosophical folk psychology.

Having this analysed *M* (or "belief", for example) according to the causal role defined by its position in folk psychology, Lewis can achieve a **token identity** account of the role of mental states such as beliefs and desires in the production of action:

> Mental state *M* = the occupant of the *M*-role (by analysis),
> Physical state *P* = the occupant of the *M*-role (by science),
> Therefore *M=P*
>
> (Lewis 2006: 59).

Lewis argues that the *causal roles identified by his analysis continue to hold so long as the philosopher of folk psychology correctly assigns the proper mental state to its appropriate causal role*[9].

The contentious step is to grant causal efficacy to the states identified by the terms "belief" and "desire". This is the principal source of the disputable position that is the focus of the present thesis, as articulated by Bermúdez (2005: 54):

> The generalisations of commonsense psychology are, quite simply, causal generalisations and the explanations and predictions offered by commonsense psychology are causal explanations that should be understood in the way that

[9] A declaration of an essentialist commitment.

> causal explanations have been classically understood – namely, as involving the subsumption of two events under a general causal law[10].

Among the most vociferous proponents of classic philosophical folk psychology is **Jerry Fodor (1975, 1987, 1999)** who argues not only that the belief-desire picture is central to questions of action choice and interpersonal understanding, but also that *the causal relationship captured by the assumptions and the belief-desire law directly reflects the functional organisation of the human brain*. **Computational processes** take place over **tokens** encoded in the brain. Representations betokening information (beliefs) combine with representations of motivations (desires) to stimulate motor responses (actions) at the neurological level. This is the essence of the **Representational Theory of Mind** (Fodor 1987; Von Eckardt 2012) that, for Fodor, describes the basis of all human cognitive processes. The presumed isomorphism between the kinds of action-description (predictions and explanation) offered by "the folk" and the computational operations of brain-circuits gives rise to the theory that neural representations are related to one another in *sentence-like combinations* in a **Language of Thought** (Fodor 1975, 2008). This suggests that the computational processes of the brain manipulate symbols that encode the ideas, perceptions and other cognitive contents in systematic ways to generate outputs, including actions. In common with natural language, the relationship between these signs and what they signify is arbitrary[11].

Fodor is scathing in his rejection of any suggestion that the belief-desire picture might not be true. He once wrote that should the belief-desire picture prove to be erroneous it would be "…beyond comparison the greatest intellectual catastrophe in the history of our species" (Fodor 1987: xii). In more temperate language he describes the relation between beliefs and desires and action choice thus:

[10] In Chapter 1, Section 1.1, I question whether *subsumption under laws* accurately describes the project of psychology.

[11] In contrast to the *embodied* or *enactive* approaches that we will encounter in Chapter 7.

> The natural home of the propositional attitudes is in 'common sense' (or 'belief/desire') psychological explanation. If you ask the Man on the Clapham Omnibus what precisely he is doing there, he will tell you a story along the following lines: 'I wanted to get home (to work/to Auntie's) and I have reason to believe that there, or somewhere near there, is where this omnibus is going'.
>
> Fodor (1991a: 23)

This restates the central commitment of the belief-desire law. *Fodor maintains that that this kind of reason-giving is a variety of causal explanation* (Ibid.). His first justification for this commitment is to point out that counterfactuals are readily available: had the Man not believed that the Clapham Omnibus was headed to his desired destination, *he would not have been on board at all*. More directly, the causal efficacy of beliefs and desires is mandated because their tokens (to get home, to work or to his Auntie's) would be **instantiated** in the computational machinery of his brain. Thereafter, it is a straightforward matter of empirical anatomy to identify a mechanism[12] whereby these tokenings play a part in neurochemical processes leading to muscle contractions and so to movements. Fodor (1991a) describes himself as a **realist** about mental representations (tokened as neurological states) and maintains that the isomorphism between propositional attitudes and their **instantiating brain states** permits him to be a *realist about the causal effects of propositional attitudes*.

Perhaps the *most influential account of folk psychology* is that offered by **Donald Davidson (2001).** Alvarez (2010), having first noted the significance of these questions of agency throughout the history of Western philosophy, writes:

> … Donald Davidson's work from the 1960s and 1970s is seminal. In particular, Davidson's conception of reasons, or something close to it, became the orthodoxy and remains so to this day. Following Davidson, most philosophers today maintain that a person's reason for acting is a combination of a belief and a desire.

Davidson's approach, known as "anomalous monism" (Bermúdez 2005: 45; Davidson 2001: 214), depends on the observation that although the physical instantiation of the mental states "belief" and "desire" (typically) is the sole location of causal effectiveness (hence

[12] An entirely *physical* mechanism.

monism), mental states give rise to one another: each is caused by antecedent states. Davidson's monism is anomalous in that it permits mental states to have causal effects. We build our understanding of mental states, Davidson argues, from this network of causal relationships, augmented by notions of *rationality* (with the belief-desire law at its core), *consistency* and *coherence*. Bermúdez (2005: 43) describes Davidson's view: "The process of interpretation is essentially a process of rational reconstruction aiming to maximise the rationality of the agent whose behaviour is being reconstructed." We presume that agents act consistently with their true beliefs and a set of goals or desires and are justified in understanding that their actions result from the relationship between them. Davidson writes:

> Any effort at increasing the accuracy and power 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 these systems from the evidence, we necessarily impose conditions of coherence, rationality and consistency.
>
> (Davidson 2001: 231)

Thus a conception of what it is to be normatively rational determines an understanding of the roles beliefs and desires play in our mental lives. In contrast, scientific explanations at the sub-personal level are *descriptive* rather than *normative* (Bermúdez 2005: 43) and so construct claims of a different kind. Davidson goes on:

> These conditions have no echo in physical theory, which is why we can look for no more than rough correlations between psychological and physical phenomena.

In the very next paragraph, Davidson establishes a manifesto for philosophical folk psychology:

> Consider our common-sense scheme for describing and explaining actions. […] we can explain why someone acted as he did by mentioning a desire, value, purpose, goal or aim the person had, and a belief connecting the desire with the action to be explained.
>
> (Ibid.)

It is through Davidson that so much contemporary philosophy and the other disciplines that depend on the philosophical view for the way that they express the concepts of action and

agency, have largely accepted the notion that *a reason, when described in a particular way, can also designate the cause of an action*. This can be done, the Davidsonian philosopher would maintain without embracing **dualism**, so long as we keep in mind the distinction between the normative requirements of rationality and the descriptive objectives of scientific enquiry.

The worry that Davidson's "reasons" would have to be instantiated in a physical system isomorphic with the grammar of reasons – part of the basis for Fodor's Language of Thought Hypothesis (Fodor 2008) – motivates an alternative view of the relation between reasons and causes, at least with respect to interpersonal understanding. The **intentional stance**, proposed by **Dennett (1989)**, attempts to sidestep causal worries by remaining agnostic about the relationship between beliefs and desires and the causes of behaviour. Dennett suggests that negotiating the physical and social world while maintaining our individual sense of agency is a matter of *taking the stance that we, and others, act in ways that are motivated by our desires and informed by our beliefs*.

> The *intentional stance,* in which *one treats the system whose behaviour is to be predicted as a rational agent*; one attributes to the system the beliefs and desires it *ought* to have, given its place in the world and its purpose, and then predicts that it will act to further its goals in the light of its beliefs.
>
> (Dennett 1988 emphasis added)

This noncommittal stance over whether the entities of the belief-desire law pick out the *causes of behaviour* – raises Fodor's question: *if it's not true, how does it work?* Dennett suggests one way of understanding how the intentional stance works:

> Evolution has designed human beings to be rational, to believe what they ought to believe and want what they ought to want. The fact that we are the products of a long and demanding evolutionary process guarantees that using the intentional strategy on us is a safe bet.
>
> (Dennett 1997: 76)

Dennett concedes that this explanation is "uninformative". He admits that there is *no known description of the mechanism by which the intentional strategy is supposed to work*. It is

notable that he brings to bear similar *normative considerations* – two occurrences of the word "*ought*" – to those suggested by Davidson.

The intentional stance is a description of the role that beliefs and desires are presumed to play in interpersonal – and even **intrapersonal**[13] – understanding. Regardless of whether the individual taking the stance is *correct* when they attribute specific beliefs and desires to an actor in identifying the causal antecedents of the act to be described or explained, it is presumed that such attributions are the basis of our appreciation of actions.

Despite its neutrality on the causal efficacy of beliefs and desires, Dennett's account respects the assumption that puts the "folk" in "folk psychology": that the attribution of beliefs and desires is a universal – or near universal – strategy in the understanding and explanation of action. Further to this, *reason-giving* is presumed both to be an example of this kind of explanation and to consist, primarily, of the introspective self-attribution of beliefs and desires.

Computational accounts of the workings of the human mind are often collected together under the heading of **functionalism** (Armstrong 1988; Block 2004; Putnam 2002).

According to Kim (2011: 129), functionalism (specifically **machine functionalism**) superseded **[type] identity physicalism** – the notion that mental states were to be identified with the brain states that instantiated them – within a few years of the reception of Hilary Putnam's paper "*The Nature of Mental States*[14]" (Putnam 1980). Taking the example of *pain*, Putnam argued that if the state were to be identified with a specific neurological occurrence, this would preclude the recognition that a creature with markedly different physiology from that of a human could experience pain. If "pain" signified the discharge of a particular set of neurons in a human brain, then we could not say that a creature that lacked that set of neurons experienced pain, however much its behaviour in the presence of physical harm resembled a pain response, without a change of meaning to the term "pain". On the other hand, we would not want a mental-state term like "pain" to collapse into a set of **behavioural dispositions** as postulated by the **logical behaviourism** associated with Ryle (1949, 2000).

[13] Self-awareness.

[14] Originally published in 1967 as "Psychological Predicates" in Capitan and Merrill (eds.) *Art, Mind and Religion* (Pittsburgh PA: University of Pittsburgh Press)

Instead, Putnam suggests that mental states, including pain, are identified with the "state of receiving sensory inputs which play a certain role in the Functional Organisation of the organism." (Putnam 1980: 229). Specifically, the role of referred to by "Pain":

> Is characterized, at least partially, by the fact that the sense organs responsible for the inputs in question are organs whose function is to detect damage to the body, or dangerous extremes of temperature, pressure etc., and by the fact that the "inputs" themselves, whatever their physical realization, represent a condition that the organism assigns a high disvalue to.
>
> (Ibid.)

This, then, is a partial functional definition of "pain". Other mental states can be described according to other functional roles.

> According to functionalism, a mental kind is a functional kind, or a causal-functional kind, since the "function" involved is to fill a certain causal role.
>
> (Kim 2011: 133)

Here we can see the origins in a functionalist conception of mind of both the belief-desire law and the metaphysical commitments that specific beliefs and desires must combine to bring about specific actions. The mental state designated by "belief", in the context of the belief-desire law, is *that mental state that causes a person to act in a particular way in pursuit of an associated desire*. That is its functional definition. As Stich and Nichols (2003: 238) argue, functionalism and philosophical folk psychology are inseparable because "According to functionalism, *folk psychology is the theory that gives ordinary mental state terms their meaning*." (emphasis in original).

Functionalists seek to explain how we can describe mental activity independently of the instantiating material without positing any additional substance of which minds are constructed. One way to describe this is that *minds are what brains do* (in the case of human minds). **Functional states** may be considered autonomous from the material of the neural apparatus that instantiates them to the extent that they might be **multiply realisable**, which is to say identical functions could be performed by quite different physical systems – perhaps even by man-made machinery. Functionalism allows the description of human mind-events to be divided into two sets of facts (Godfrey-Smith 2004); those concerning

the functions and those concerning the brain-events that are their instantiation. This does not entail separation into two substances. Functionalists do not subscribe to the idea that functional mental states are identified with any kind of *spooky non-physical, non-spatial "stuff"* that (somehow) provides the impetus for action.

All of this is perfectly reasonable – except that for functionalism to be useful it is *essential that our models correctly identify the functional states*. Central to much *philosophical functionalism* – which Bermúdez (2005: 58-61) differentiates from *psychological functionalism* – is an assumption that the belief-desire law is an accurate description of the functional (causal) roles of the entities that make it up[15]. In Chapter 1, I describe the process of **functional analysis** by which scientific psychology seeks to *(causally) explain complex functions in terms of the simpler functions that systematically combine to perform them*. My contention is that if the functional roles identified by the belief-desire law and the metaphysical commitments were performed by mental states answering to "belief" and "desire" then we would find these terms, or the states that they pick out, featuring either as the objects of functional analysis or at some stage in that process. Otherwise, *the assumption of the functional role of belief and desire is unwarranted*.

This thesis does not advocate a return to behaviourism. In common with the cognitive and social psychology that I turn to in Chapters 2 and 3, I urge philosophers to be *intensely interested in the functional processes that underlie behaviour*. Behaviourists typically treat the realm of the mental as a "black box" whose contents are both impenetrable and irrelevant to the investigation of relationships between stimulus and response (Kim 2011: 84; Skinner 1974: 61). I am content that mental states can be differentiated by their functions but *question whether "belief" and "desire" clearly designate unique functional states, and, consequently, whether the identification of these terms with particular functions is warranted*.

0.3 Philosophical Problems Generated by Folk Psychology

One of the most frequently rehearsed problems with the belief-desire picture is the **problem of mental causation**. As Broome (2013), in the context of the normative prescription of behaviour, puts it:

[15] It might be conceded that the neural tokens are the actual "cause" of the action or action-choice but this makes no difference to the presumed isomorphism between functional and causal roles.

> …when you believe you ought to do something your belief often causes you actually to do it. … One part of the mind-body problem is to understand how a state of mind can have a physical effect like that.
>
> (Broome 2013: 1)

Although the mind-body problem is engendered by the belief-desire picture, it will *not* be given any direct consideration in this thesis. It might be completely intractable, in common with many metaphysical questions. It might also be something of a **pseudo-problem** in that it emerges only if one assumes that the belief-desire law is binding, that "belief" and "desire" (as they appear in that law) pick out the causes of action.

A structural weakness of the belief-desire picture emerges from one of its strengths. The belief-desire law could, conceivably, explain any and every possible action, regardless of whether or not the individual subject acts in accordance with their professed beliefs and desires. All one needs do to explain anomalous behaviour in this way is to posit a set of unspoken beliefs and desires, perhaps even an **unconscious** set of beliefs and desires. So even when somebody finds it hard to account for their own behaviour in belief-desire terms, it remains possible to preserve the essential causal efficacy of beliefs and desires – the actor is just unaware of them.

One could imagine an infinite variety of beliefs and desires that could be suggested – on the assumption that the belief-desire law pertains – to "explain" any action, regardless of whether the actor would recognise these ascribed mental states as their own. People might not be all that reliable about their own beliefs and desires, after all[16]. This is the basis of much discussion of **akrasia** or "weakness of the will" as it is sometimes described (Davidson 2001: 21-42). A smoker might openly declare their *desire* to quit and their *belief* that smoking presents a serious risk to their future health and longevity (and presumably also desires to live a long and healthy life). And yet they continue with their habit. This leads Wilkes (1991), for example, to argue that philosophical folk psychology, unlike scientific psychology, is *not equipped* to deal with "irrational or non-rational behaviour" because the models it develops are **normative**; they prescribe *what the ideal decider ought to decide under ideal decision-conditions*. Such conditions would, presumably, be the

[16] Which immediately raises doubts about the value of reason-giving on the belief-desire model.

absence of conflicting bodily sensations – such as the addictive physical effects of nicotine – or such psychological effects as **hyperbolic discounting** (Laibson 1997) under which long-term risks are discounted in favour of short term gains even when those risks (painful sickness and premature death) enormously outweigh any benefits (a warm, fuzzy feeling of drawing drug-laden smoke into the lungs).

Conflicts between action and expressed beliefs and desires are resolvable. We just have to tell the unwitting actor that deep down (on some level, subconsciously etc.), they *must have believed that what they did would get them what they (also unconsciously) wanted*. The belief-desire law and the assumptions remain intact. This weakens the explanatory value of the belief-desire picture because it renders the strategy **unfalsifiable** (Popper 1992). What explains everything in this way, in fact explains nothing since we can always postulate additional contingencies that fit the actuality. It would be impossible to distinguish, by examination of any specific belief-desire pair that is suggested as an explanation, between any genuine causal antecedents of the observed action and a plausible **rationalisation**.

There is also considerable debate in the philosophy of action as to whether any law-like generalisation can accommodate ceteris paribus clauses: can we regard any statement as a law if the force of that law is contingent on unspoken conditions or exceptions? See Fodor (1991b) and (Gauker 2003) for alternative views.

A practical difficulty arises from the presumed role of belief-desire psychology in interpersonal understanding. In supporting a version of FP, Gopnik and Seiver (2009) claim that:

> Negotiating the social world is an extraordinarily difficult and complex task.

Would we want to accept this? I suspect that behind Gopnik and Seiver's statement is the thought that *explaining* how individuals negotiate the social world is "difficult and complex"[17]. For most people – those not on the more socially disabling parts of the autistic spectrum, for example – most everyday social transactions and interactions are straightforward, automatic[18] and even enjoyable. Empirical evidence suggests that a lack of interpersonal contact can be seriously detrimental to psychological and even to physical health (House et al. 1988). Far from being taxing or requiring the application of an algorithm

[17] See Chapter 3.

[18] See Chapter 3, Section 3.6 for a discussion of the automaticity of social cognition.

based on an ability correctly to ascribe mental states – such as beliefs and desires – to those around us, many social inferences seem *effortless* (Bargh and Ferguson 2000). Contrast this with the frequent description of social skills as "mind-reading" (Apperly 2011; Baron-Cohen 1995; Butterfill 2013; Carruthers 2009; Currie and Sterelny 2006; Goldman 2006). The belief-desire picture generates an appearance of complexity where, for most people and for most of the time, there is none[19].

I have already mentioned the dispute between different conceptions of how folk-psychological proficiency is grounded – between simulation theorists and theory-theorists (and others). Philosophical folk psychology is reliant on empirical evidence for this dispute. It cannot be resolved by conceptual means alone. Chapter 3 of the present thesis, however, will show some ways in which scientific psychology accounts for the capacity to predict and explain one another's actions *without recourse to guessing, inferring, simulating, or otherwise ascribing sets of beliefs and desires* to agents.

Even without attempting to account for *anomalous behaviour* – those instances when an individual acts in a way that is at odds with their declared beliefs – there is a need to accommodate individual differences and choices. Take the way that somebody is understood, on the belief-desire model, to take a political stand. Suppose two people both believe, wholeheartedly, the proposition "slavery is wrong". Person A declares their vehement opposition to slavery and their support for measures to eradicate slave ownership. Person B does the same but, additionally, attends protests against the practice. Although the moral realist would say that some feature of slavery compels person B to protest, whereas the anti-realist would account for their decision to protest by the possession of a *belief* in the wrongness of slavery, both would say that the belief in the proposition "slavery is wrong" is among the causes of person B's decision to march.

Given that persons A and B share the same attitude to the proposition "slavery is wrong" – they believe that it is true – then the belief-desire model has to account for the difference in their behaviour. One might try to do this by positing additional, differentiating beliefs or desires (e.g., in the effectiveness of demonstrations). It remains possible, however, that two people could share identical attitudes and yet act quite differently. Belief-desire psychology

[19] This is not to downplay everyday instances of social anxiety: however, these tend not to be generated by a failure to appreciate the intentions of other people, but rather by concern over how one's own actions will be perceived (Schlenker & Leary 1982).

makes no attempt to account for complex **affective responses** (Zamuner 2015) to a proposition (how one *feels*, for example, about the wrongness of slavery) that can underlie these **individual differences**, without resorting to a proliferation of causally efficacious beliefs and desires[20].

At the other end of the spectrum, belief-desire psychology easily accounts for the behaviour of a housefly. Stimulated by chemical signatures given off by food, the fly is directed by the differential strength of chemical stimuli between its two antennae to find the source, settle and feed. Each turn that it makes is determined on the basis that it *desires to feed* and (thanks to its sensory information) *believes that the food lies in a particular direction*. Humans are, according to philosophical folk psychology, just more complicated versions of the fly: our beliefs and desires may be more complex, layered and even conflicting. But they determine, ultimately, what action we will take.

A contention of this thesis is that *people are both more complicated than flies (i.e., subject to many more kinds of influence) and yet not so complicated that we need an almost infinite nested series of beliefs and desires to account for what they do*.

One response to the problems generated by philosophical folk psychology is **eliminative materialism**. For more than thirty years this position has most closely been associated with Churchland (1981), although its fundamental features can be found in Feyerabend (1963) and Rorty (1970). The eliminative materialist suggests that philosophical folk psychology attributes a theory to people, but that, in common with "folk physics" and "folk biology", advances in science are likely to prove the folk theory wrong. The appeal of folk psychology to philosophers of mind will consequently be undermined by new and better theories in neuroscience. Indeed, the theoretic entities of the folk theory of mind – beliefs and desires – are likely to be *eliminated and supplanted by the more explanatorily complete and predictively successful elements of the new theories*, such as neurochemical events or electrical firings at the level of the individual neuron. Within a few generations, Churchland's 1981 paper suggests, even the everyday "folk" will have no need of the language of "belief" and "desire" to compose explanations of their own and others' actions any more than they currently use the language of *crystal spheres*, *vital spirit* or *phlogiston* in their respective domains.

[20] And a further need to account for the order of precedence between conflicting attitudes.

Whereas the eliminative materialist might contend that there is **no such thing as belief**[21], – i.e., the word "belief" has no actual referent (Garfield 1988: 4-5; Stich 1983) – in this thesis I argue that because "belief" and "desire" do not necessarily pick out the functional roles suggested by philosophical folk psychology, philosophical problems are generated when we unwittingly use them as if they do. Further, I will demonstrate that there are a *great many more things* (concepts, functions, explanatory roles, demonstrations, defences etcetera) *that answer to "belief" or "desire" than philosophical folk psychology can accommodate*. Confusion arises when philosophers, and non-philosophers who rely on the clarity expected of philosophical discourse:

> ...overlook that the terms 'belief' and 'desire', as they usually feature in the folk psychology debate, turn out to be placeholders that encompass a wide range of psychological predicaments which everyday 'folk' routinely distinguish with ease.
>
> (Ratcliffe 2008)

The tendency to use "beliefs" as if the word were defined solely by the causal role of its presumed referents crops up in **Antony (2015).** In a defence of the *truth of folk psychology* she offers an exposition of Simon Blackburn's[22] reconstruction of an argument in support of eliminative materialism that both find in Stich (1983):

1) Functionalism a) defines mental states by their causal role vis-a-vis inputs, outputs and other mental states and b) draws these definitions by idealising from the platitudes of our everyday practices of mentalistic ascription. Moreover, c) functionalism entails that the functional organisation so derived must be "realised" in whatever kind of matter composes the psychological being in question.

2) For human beings, psychology is realised in neurophysiology.

3) According to the most natural idealisation of our practices of belief ascription, a belief is the kind of mental state that can be causally responsible for both verbal behaviour and non-verbal motor behaviour.

[21] This is, to some extent, a parody, often presented as a *reductio ad absurdum* argument against eliminative materialism.

[22] Anthony's chapter is taken from a collection celebrating the philosophy of Simon Blackburn.

4) Therefore, for human beings, there must be, for every belief, a single neurophysiological state that plays the role of causing both verbal and non-verbal behaviour [from (1c), (2) and (3)].

5) But there is no such neurophysiological state [empirical evidence]

6) Therefore, no human beings possess any beliefs.

Anthony agrees with Blackburn that Stich's eliminative materialism, as reconstructed here, rests on a non sequitur in the move from 3 to 4. It might not be the case that functionalism is committed to view that each functional state must be differentiated at a one-to-one ratio into realising neurophysiological states. Her dispute with Blackburn rests on the notion that he "concedes too much to the eliminativist; he doesn't challenge the reductivist assumption that the empirical test of psychology will lie in the domain of neuroscience," (Antony 2015: 17).

Note however that neither Anthony, nor (on her reading) Blackburn, challenges the definition of belief enshrined in step 3. Drawn from the assumptions of belief-desire psychology, this is taken as a given. The challenge to the eliminative materialist – like Stich's eliminativism itself – depends on moves that come after this clause is presented without contest.

Broome (2013) is explicitly committed to the causal efficacy of propositional attitudes although, as mentioned above, he is aware of the metaphysical "worries" this engenders. His work is primarily concerned with a description of the role that rationality plays in moral decision making, and the role of the **enkratic principle** – that a person should *intend* to behave in the way that they *believe* that they ought to behave. A belief in specific normative requirements thus becomes a component in the formulation of an intention to act. The enkratic principle is distinct from the belief-desire law, although clearly related: not least in the functional role that it ascribes to beliefs. Although Broome announces at the outset that *he will set the mind-body problem aside* by "…focussing on your intention rather than your action. The motivation question is about your mind only" (Broome 2013: 1) he is compelled, in the very next sentence, to assert that:

> When your belief causes you to intend to act, your intention will in turn generally cause you to act, but that is not my concern.

Broome thus acknowledges that the causal generalisation gives rise to problems – in this case the mind-body problem. Nonetheless, he prefers to leave belief-desire psychology intact and work instead on normative aspects of action choice. In the background, the unwarranted assumptions remain.

Thornton (2009) considers the **interface problem**, as coined by Bermúdez (2005):

> How does commonsense psychological explanation[23] interface with the explanations of cognition and mental operations given by scientific psychology, cognitive science, cognitive neuroscience and the other levels in the explanatory hierarchy?
>
> (Bermúdez 2005: 35)

In order to regard psychological explanations as *autonomous* while at the same time being committed to physicalism/materialism there must be some point at which explanations at the personal level interface with those at the instantiating sub-personal level – assuming, of course, that these "explanations" are equally true and equivalent in terms of explanatory and predictive utility.

Thornton's approach to this is to "question the metaphysical pretentions of nomological science" (Thornton 2009: 121 Abstract). The non-nomological nature of psychological explanation will be examined in Chapter 1 of the present thesis. However, in trying to localise psychological explanation, Thornton is committed to the view that it is a real issue, arising from the human agent's *dual nature* as both a physical and an intentional system. He touches on Dennett's intentional stance, Davidson's anomalous monism and Fodor's representationalism. He also draws on the distinction between "manifest" and "scientific" images suggested by Sellars (1962) as the origin of this concern. His conclusion is that we should call off the search for a global reductive strategy – and thus universal solution to the interface problem – in favour of a series of "local" solutions. He writes:

> Provided the apparent need for a global account of the interface between higher and lower levels can be eased, there is space for local accounts of how

[23] "Common-sense psychology" (as encountered in Davidson, above) is preferred by some scholars on the grounds that it carries none of the pejorative connotations associated with the term "folk psychology".

> descriptions at the level of the whole person can interact with underlying cognitive neuroscience.
>
> (Thornton 2009: 135)

Campbell (2009) questions whether the *normative standard* implied in the belief-desire law is of use in determining whether an individual is in need of psychotherapeutic intervention[24]. In doing so, he tackles mental causation and suggests that we can regard propositional attitudes either as mechanistic causes of behaviour or as *control variables* – entities whose presence *might not causally entail the performance of an action but whose value might make it more or less likely to occur*. Whether or not a person is sensitive to the control of their propositional attitudes is, he suggests, a way of differentiating pathologies of behaviour without recourse to a normative ideal of rationality:

> If the propositional attitudes function as control variables in this sense, then we do have a causally functioning mental life, whether or not the subject is rational. Of course, it is true that in the mental life of a broadly rational subject, propositional attitudes function as control variables. But it is the fact that we have control variables, not the fact that we have rationality, which means that we are 'at the right level' to talk of beliefs and desires.
>
> (Campbell 2009: 147)

Note that for Campbell, whatever way we choose to describe the mode through which they bring about their effects, the notion that *propositional attitude terms pick out causes of behaviour* is not questioned

A more sceptical account that is nonetheless a direct reaction to this prevalent philosophical view of the relationship between beliefs and action choice is offered by **Bortolotti (2010)**. In seeking to establish that *delusions are a variety of belief,*[25] she begins by engaging with the contemporary philosophical literature on the subject in order to establish what features qualify a given mental state as a belief. In doing so, she touches on an account of rationality that has at its heart the role of "belief". For instance:

[24] The notion being that somebody who consistently acts at odds with their own beliefs and desires is in need of an intervention.

[25] Which, she points out, is disputed on normative grounds.

> …I behave irrationally because the belief-like state I have reported fails to be consistent with my subsequent behaviour.
>
> (Bortolotti 2010: 10)

Despite this observation of the *normative* picture of the relationship between the possession of a given belief and the performance of an action she is sceptical that a satisfactory philosophical definition of "belief" is available. For instance:

> Delusions help me make salient and relevant the observation that the states we ascribe to ourselves and that we call 'beliefs' are very heterogeneous. They can have some typical belief-like features (such as being manifested in action or being used in inferences) without satisfying norms of rationality that some philosophers have regarded as preconditions for mentality, and more specifically, for the possession of beliefs.
>
> (Bortolotti 2010: 3)

and

> Two things should be immediately noted about the proliferation of rationality constraints. First, one common point to all variations is that rationality is supposed to be a necessary condition for belief ascription. If there is no rationality, then belief ascription is impossible or illegitimate. Second, the implications of the view will vary according to the following factors: how rationality is defined, and to what extent the subject's behaviour or belief-like state has to diverge from standards of rationality in order for the ascription of beliefs to be impossible or illegitimate. …
>
> It is a real challenge to provide a definition of what beliefs are, let alone an account of the necessary and sufficient conditions for believing that something is the case.
>
> (Bortolotti 2010: 11)

Although Bortolotti may be sceptical about the role of beliefs as action-causing or action-guiding mental states the problem that she is reacting to is of a sort with those emerging from the philosophical folk psychology picture. Delusions, so the thought goes, cannot be

regarded as beliefs because they fail to exhibit the normatively "rational" function of regulating behaviour. She writes

> Notice that the degree to which the reported state is, in fact, a belief does not indicate the subject's level of confidence in the believed state of affairs, but the extent to which her behaviour can be legitimately characterised by the description of beliefs.
>
> (Bortolotti 2010: 20-21)

It should be noted that the "problem" whereby "delusions" cannot be regarded as beliefs – even as incorrect beliefs – emerges from the picture of beliefs as being **essential** to a model of rationality enshrined in philosophical folk psychology. Bortolotti's response is a traditional analytic philosophy move to alter the definition both of "belief" and of "rational" so that delusions can be encompassed under their umbrella. Responses and objections to Bortolotti's thesis have been on similar grounds – usually objecting that her extension of definitions is unwarranted or unwelcome; see, for example, Tumulty (2012).

0.4 Philosophical Pictures, Dogma and the Therapeutic Approach

Many of the problems arising from philosophical folk psychology arise from its status as the default, dogmatic **picture** of action, interpersonal understanding and reason-giving. The fixed viewpoint generated by this picture is an obstacle to a "clear view" or a "perspicuous representation" (Baker 2004: 182-83) of the philosophical questions, in the sense highlighted by Wittgenstein in the *Philosophical Investigations*, §122. Instead, philosophers construct an image of the issue that Gordon Baker (2004: 32-33) characterises as "continuous aspect seeing". From this position, any appreciation of alternative ways of tackling the issue is precluded and problems are generated as artefacts of a point of view rather than being features of the matter under consideration. The philosopher who sees this picture from their continuous aspect would be unaware of the limitations imposed by their rigid and **essentialist** position. As Morris (2004: 7) puts it; "The person behaves intellectually as if his picture represented *the only possibility*."

One species of such difficulties, arising from the use of analogies and metaphors, is described by Fischer (2011b: 21):

> We are *under the spell of a philosophical picture* when our philosophical reflection is guided by certain analogies within language, without our being aware of being guided by them

On this reading, the relevant intuitive conclusions are spontaneously inferred through analogical inferences with conceptual metaphors. (Fischer 2011b: 22-28). These analogies can become so deeply entrenched that the philosopher fails to notice them as such. This will mean that problems or puzzles are generated because features of the model *source domain of the conceptual metaphor are unwittingly projected onto the target domain*. This puts the philosopher at risk of deriving inferences through *non-intentional reasoning* (Fischer 2011b: 28-35), in which the metaphorical/analogical implications of the picture play a significant yet unnoticed part.

For my purposes, however, I do not need to show that philosophers are misled by commitments arising from metaphors such as "the mind is a machine" or "thinking is like physical activity". I suggest only that the **philosophical picture** is a fixed, dogmatic and prejudicial (Morris 2007: 69) view of a topic such that the philosopher risks being unable to imagine any other way of seeing the matter. At the same time, just in virtue of its ubiquity and fixedness, the picture generates apparent **philosophical problems** arising from "tacit and unwarranted presuppositions at odds with warranted beliefs the philosophers raising the problems reflectively hold at the same time" (Fischer 2006).

In the case of the belief-desire law and the assumptions of philosophical folk psychology the risk of being in the sway of a problem-generating philosophical picture arises not from *unwitting analogical inferences* but from *unwarranted theorising*:

> A significant number of philosophical problems are being raised only due to some 'implicit theorising' or, more accurately, due to drifts of thought in which we tacitly presuppose substantive philosophical assumptions, without realising it.
>
> (Fischer 2006)

Wittgenstein frequently points out the dangers of being under the sway of theories that amount to philosophical dogma (cf. Kuusela 2008):

> If I rectify a philosophical mistake and say that this is the way it has always been conceived, but this is not the way it is, I must always point out an analogy according to which one had been thinking, but which one did not recognise as an analogy.
>
> Wittgenstein, *Big Typescript* §408

In more developed form in the *Philosophical Investigations*, he writes:

> For we can avoid unfairness or vacuity in our assertions only by presenting the model as what it is, as an object of comparison – as a sort of yardstick; not as a preconception to which reality must correspond (The dogmatism into which we fall so easily in doing philosophy.)
>
> Wittgenstein, *Philosophical Investigations* §131

For Gordon Baker, the escape from pictures is the central goal of Wittgenstein's philosophy:

> 'Our method' tries to bring to an individual's consciousness the influence of pictures working unconsciously within him. It strives to combat pictures that generate perplexities or confusions.
>
> Baker (2004: 185)

"Our method" refers to the **diagnostic-therapeutic** approach that Baker reconstructs from his readings of Wittgenstein and of Waismann (1968). My contention, after Baker, is that unwarranted commitments give rise to philosophical problems associated with the belief-desire picture. Philosophers are motivated to address these problems by advancing further theories as potential solutions. The objective of the diagnostic-therapeutic method in this thesis is to suggest that *a wider view of the phenomena can liberate philosophers from the background assumptions that generate the problems and give rise to unwarranted philosophical theorising* (Baker 2004 passim). The goal is "either eliminating these objects or altering someone's attitude to them" (Baker 2004: 183). "These objects" in this context being *"belief" and "desire" as they appear in causal accounts of action*.

Why should the goal be the elimination of "belief" and "desire" rather than simply the modification of philosophical folk psychology in order to accommodate the evidence? Many of the philosophical problems generated by folk psychology – including those

pursued by the contemporary philosophers in the last section – are generated by the pursuit of just such an accommodation. *Making a generalised causal claim frequently entails setting ordinary terms to metaphysical uses.* To avoid perpetuating the dogma, its resulting problems and further intellectual disquiet, philosophers should exclude "belief" and "desire" from generalisations about the causes of action. This is the "**modest eliminativism**" to which the present thesis is directed.

According to Baker (2004: 182) this objective for philosophy pervades Wittgenstein's writings.

> When philosophers use a word – "knowledge", "being", "object", "I", "proposition/sentence", "name" – and try to grasp the essence of the thing, one must always ask oneself: is the word actually ever used in this way in the language in which it is at home?
>
> What *we* do is to bring words back from their metaphysical to their everyday use.
>
> Wittgenstein: *Philosophical Investigations* §116

In the present thesis I set out to bring "belief" and "desire" back from their metaphysical to their everyday uses.

Philosophical folk psychology commits the philosopher to the claim that "belief" and "desire" pick out the *causes* of actions. The metaphysical and the everyday meaning have parted company. More damagingly, they have done so under the philosopher's nose. Proponents of philosophical folk psychology start from an account of what they assume to be the everyday, *non-philosophical*, understanding of agents and observers ("the folk") before going on to argue for ways that the assumptions they attribute to the "folk" are (more or less) correct. Much of the persistence of belief-desire psychology stems from *the assumption that the familiar use is automatically being respected* – this is, after all "folk" psychology. Hence the unwarranted commitments that the view entails blend into the background.

The diagnostic-therapeutic method begins with paying attention to what *actually happens*, both in terms of the target phenomena (action choice – interpersonal understanding – reason-giving) and the working uses of the ordinary terms (belief and desire) that are

appropriated for the construction of philosophical ideas and which acquire unwarranted metaphysical commitments in the process.

This approach *need not preclude the use of argument to establish the foundation from which the therapeutic project proceeds*. Fischer refers to therapeutic approaches in psychotherapy to point out that:

> …exemplary psychotherapies like cognitive therapy (J. S. Beck 1995) and rational emotive therapy (Ellis 1994) for depression crucially involve argument and the assessment of evidence (to establish, for instance, whether the depressed patient really is as utterly inept as he thinks or places an unduly biased interpretation on his own achievements).
>
> (Fischer 2011a)

Argument can bring facts to the attention of a philosopher in the grip of a picture. For example:

a) Alternative ways of viewing the phenomena exist.
b) Those alternative views are relevant to their concerns.
c) Those alternatives are free from the unwarranted commitments that generate philosophical problems.

The role of philosophical therapy in this context is as described by Baker:

> The point is [to] persuade the metaphysician to clarify precisely why he is not content to stick to this familiar use in this particular context, i.e. on why he feels driven to say something different.
>
> (Baker 2004: 103)

Horwich (2012: 11) inadvertently but succinctly summarises the therapeutic objective of the present thesis:

> The remedy, quite clearly, is not to be mesmerised by the *word*, but to appreciate how distinct uses of it, hence somewhat distinct meanings, may evolve and proliferate.

Argument in the context of diagnostic-therapeutic philosophy does not set out to replace one theory, however unwarranted its assumptions, with another and better theory analogous to a scientific **paradigm shift** (Kuhn 1996). The elimination of the title does not entail, for example, the wholesale replacement of philosophical folk psychology with a theory from neuroscience,[26] as does Churchland (1981). Wittgenstein eschewed any conception of philosophy as a process of theory development:

> And we may not advance any kind of theory. There must not be anything hypothetical in our considerations. All explanation must disappear, and description alone must take its place.
>
> Wittgenstein, *Philosophical Investigations* §109

> If someone were to advance theses in philosophy, it would never be possible to debate them, because everyone would agree to them.
>
> Wittgenstein, Philosophical Investigations §128

Neither can liberating a philosopher from a philosophical picture that generates problems be a matter of *refutation*: a picture cannot be *refuted*. It can be clarified, in order to make the philosopher in its sway aware of its implications and the degree to which their dogmatic commitments might be unwarranted; or the philosopher can be shown that alternative views of the phenomena are available, views that do not generate the same problems.

In selecting this method, I do not suggest that this is the only or even the best way to approach philosophy in general or diagnostic-therapeutic philosophy in particular. Neither would I want to maintain that Baker's reading of Wittgenstein or of Waismann is *correct*. Hacker (2007) accuses Baker of misreading Wittgenstein and overlooking Waismann's intentions. Hacker charges his erstwhile collaborator (see Baker and Hacker 2009, 2014) with failing to notice that Waismann (1968) is attempting to distance himself from Wittgenstein's conception of philosophy rather than to interpret or endorse it (Hacker 2007: 94).

Whether or not Hacker's criticism is justified and whether or not the therapeutic approach taken in these pages is one that Wittgenstein would recognise or sanction, this is the

[26] Nor even from psychology.

approach that I choose to pursue. This method is *inspired* by a reading of Wittgenstein, through Baker and (indirectly) Waismann. However, I claim no Wittgensteinian *authority* for my approach or conclusions.

Each of the five contemporary philosophers mentioned in the previous section wrestles with a problem generated by philosophical folk psychology. For Anthony, the problem is *answering the challenge from eliminative materialism*. Broome acknowledges but "sets aside" *the problem of mental causation*. Thornton recognizes that *accommodating the truth of philosophical folk psychology within scientific approaches is challenging* and concludes, *pace* Wilkes (1991)[27] that this is a problem best left to science rather than philosophy. Campbell wants *a better definition of pathological behaviour, without regard to the normativity of belief-desire psychology*. Bortolotti seeks definitions of "belief" and rationality under which *delusions can be accommodated as a class of beliefs in the face of a demand that beliefs play a part defining rational action*.

I would not suggest that any of these views are straightforwardly incorrect. All are, however, pursuing a solution to a philosophical puzzle or problem that arises from the *picture* constructed by philosophical folk psychology which is *underpinned by unwarranted commitments*. Leaving the *ceteris paribus* clause of the belief-desire law free to be populated by additional beliefs and desires, maintaining that the belief-desire law is normative rather than descriptive (Wilkes 1991), modifying the meaning of the terms in subtle ways (Davidson 2001) or restricting the scope of the commitment to *stance-taking* (Dennett 1989) each offer only partial escapes.

The therapeutic approach that I take in this thesis is distinct in an important respect from other empirically informed work with diagnostic intent. For example, building on Fischer's (2011) work on the role of intuitions in generating philosophical problems, Fischer and Engelhardt (2016) and Fischer et al. (2015) have used empirical and experimental investigations of their own, coupled with published empirical findings from psychology,[28] to develop a naturalised **metaphilosophy**. The focus of their work is on developing debunking explanations of (mainly paradoxical) **intuitions** which generate philosophical problems. The investigations that these researchers have undertaken – together with work

[27] Who would prefer that philosophers regard their version of Folk Psychology as *normative* and so not susceptible to scientific examination.

[28] Such as Alter et. al. (2007), Evans (2010) and Giora (2003).

on the **restrictionist** programme in experimental philosophy (Weinberg et al. 2012) – indicate that philosophers often have no warrant to accept intuitions that create problems, especially by clashing with their background beliefs or with each other. The only way that philosophers can inoculate themselves against the lure of pernicious intuitions is to understand how they come about.

Where the present thesis is distinct is that my engagement with cognitive and social psychology, is not intended to *explain* anything. This is also the case where I employ empirical examples from everyday discourse in part two of the thesis. My intention is to show, firstly, that an alternative way to causally explain action behaviour and interpersonal understanding is available and, secondly, that many instances of the use of "belief" and "desire" – even in the case of reason giving – do not imply causal explanations. As an *empirically-informed diagnostic-therapeutic approach, the* **therapeutic aim** of this thesis is, as suggested by Baker (2004) to guide the philosopher of mind and action to question the *prevalent philosophical picture* and to ask whether many of the philosophical problems that they seek to unravel might be features of that picture. If successful, the philosopher of mind and action who has relied on philosophical folk-psychology will be in a position to recognise the contrast between the *metaphysical and everyday* uses of terms in philosophical folk psychology as a significant source of these problems.

As seen in sections 0.2 and 0.3, above, *these commitments are widely held.* My contention is that they are held largely without warrant. We might be justified in holding these default positions even in the absence of prior grounding, so long as they do not directly contradict common sense (Williams 2001: 36). To make this contrast explicit:

> The Prior Grounding Requirement … insists that one's beliefs be *based* on adequate grounds. But there is another possibility. This is that personal justification is more like innocence in a court of law: presumptive but in need of defence in the face of contrary evidence. On this view, personal justification has what Robert Brandom[29] calls a "default and challenge" structure: entitlement to one's beliefs is the default position; but entitlement is always

[29] Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment* (Cambridge MA: Harvard University Press, 1994) at 177.

> vulnerable to undermining by evidence that one's epistemic performance is not up to par.
>
> (Williams 2001: 25)

Where empirical evidence or other directly relevant information is available that challenges the default position, the holder's entitlement to that position is weakened. In Williams' words, the default and challenge model of personal justification "replaces the Prior Grounding Requirement with a Defence commitment. Knowledgeable beliefs must be defensible, but not necessarily derived from evidence" (ibid). Nevertheless, an evidence-based challenge to one's default position can be answered only with evidence in its favour or of the reliability of the methods from which the position was derived.

To challenge the default view – philosophical folk psychology – I will firstly present evidence from cognitive psychology that action choice is not always a matter of doing whatever one believes will bring about the fulfilment of a desire. Models of interpersonal understanding (attribution theory) from social psychology will question whether the belief-desire law is the basis on which "the folk" explain and predict action. In part two I will examine uses of the terms "belief" and "desire" – including apparent reason-giving (excuses) – that do not rely on these terms referring to causal categories of mental state.

I contend that these constitute evidence that the *commitments to metaphysical uses within philosophical folk psychology* – i.e., "belief" and "desires" as picking out the causes of action – is part of the fixed way of seeing that leads philosophers, unwittingly, to misuse these expressions (Baker 2004: 94). This use is symptomatic of a philosophical prejudice. "A central feature of a philosophical prejudice is what I will call a 'perverse' attitude toward evidence and argument", according to Morris (2007: 69). This is also a core feature of *my conception* of a philosophical picture. The "perverse attitude" to evidence (especially) manifests in the philosopher seeing evidence that challenges their default position as a further philosophical problem which must be accommodated within the picture – if necessary by the formulation of a supplementary theory.

Despite my appeal to the diagnostic-therapeutic method of Wittgenstein, via Baker, I am not proposing an elimination of "belief" and "desire" from causal accounts of action on the grounds that "reasons are not causes". Greenwood (1991: 1) describes that position thus:

> A central thesis of many neo-Wittgensteinian accounts was that folk-psychological references to intentional psychological states are not causal explanatory (Louch 1966; Peters 1958). According to such accounts, a folk-psychological explanation is a logically distinct kind of explanation, one that renders them "intelligible" in the light of rules and reasons.

Greenwood's contention is that such a view would be of no interest to the scientist, who *is* concerned with constructing causal accounts (see Chapter 1).

Bermúdez (2005: 55) illustrates where this position is situated in relation to other philosophical accounts of psychology by means of the argument tree, reproduced on the next page. Following the left hand side of the tree leads the philosopher inexorably to **functionalism** (see above):

Although at first sight I might appear to take the first available fork to the right, there is a significant difference between my position and the so-called neo-Wittgensteinian view. I am making no ontological claim about the nature of reasons or causes – even if the extent of the claim is only that the two are distinct. I do, particularly in the second part of this thesis, suggest that everyday uses of "belief" and "desire" do not *entail* making a causal claim.

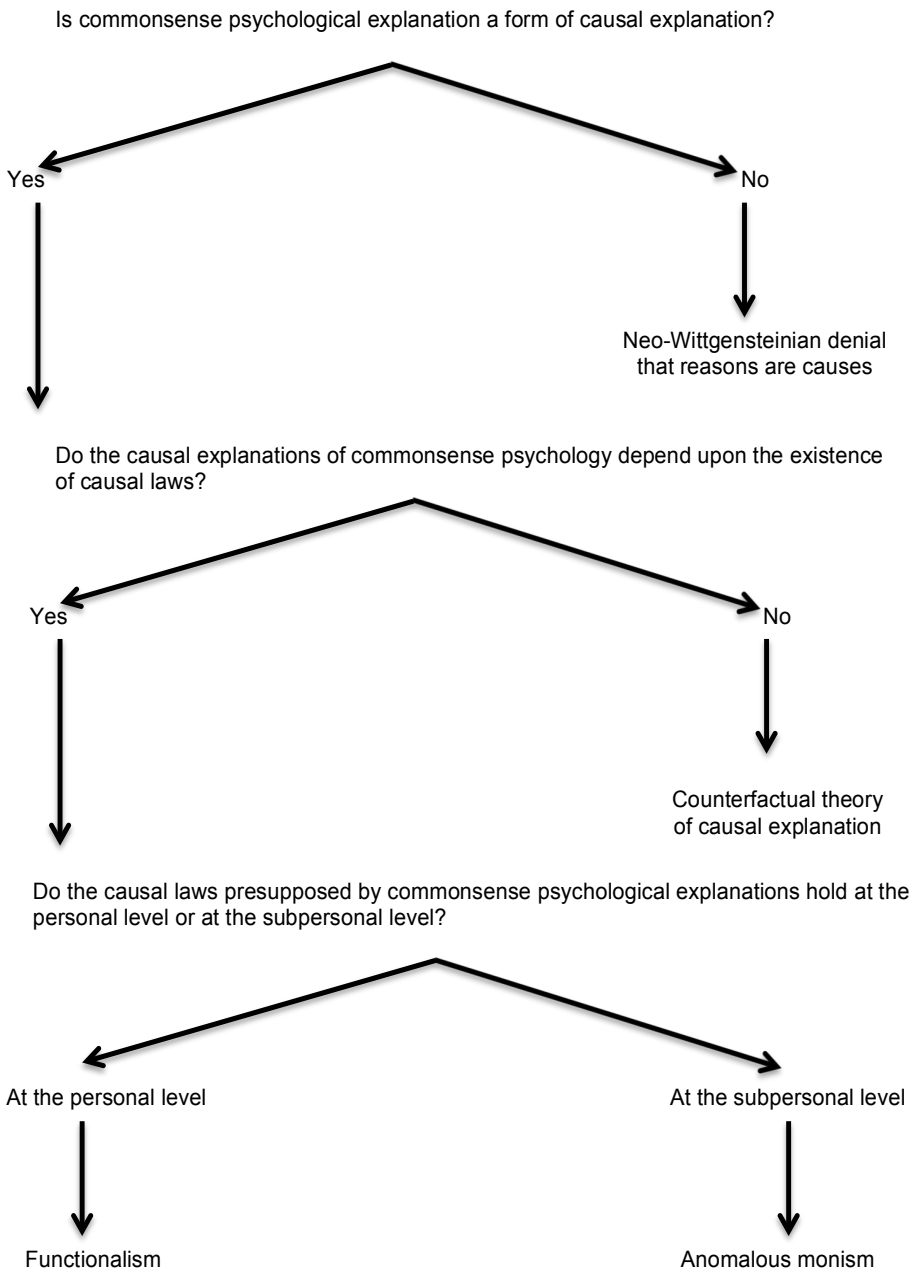

Figure 0.1: The relationship between theories in the philosophy of psychology. (Bermúdez 2005: 55).

More explicitly, *I reject the question from which the first bifurcation in Bermudez' diagram proceeds*: *I challenge whether philosophical folk psychology offers any form of "explanation", causal or otherwise.*

0.5 Synopsis of Parts and Chapters

From the two guiding questions (section 0.1) and the diagnostic-therapeutic approach (section 0.4), the present thesis divides naturally into two parts. The first will describe the approaches to action choice and interpersonal understanding taken by **cognitive psychology** and **social psychology**. The second will set out to show some of the ways in which the terms "belief" and "desire" are applied when free of the commitments of philosophical folk psychology (everyday uses, in Wittgenstein's sense).

Part One (Chapters 1-3), builds on the contention that *if* philosophical folk psychology constituted an accurate account of the *causes* behind action choice and interpersonal understanding then, given that scientific psychology is committed to providing genuine, causal explanations for psychological phenomena, we would expect to find something resembling the belief-desire law and the concepts of "belief" and "desire" at the heart of scientific explanations in action-choice and interpersonal understanding.

Chapter 1 of the thesis, immediately following this introduction, *establishes the contention that psychological sciences do seek genuine, causal explanations for its target phenomena* – including action/action choice and interpersonal understanding. On the way, I will argue that in the domain of psychology "genuine causal explanations" means something distinct from subsumption under a law (including the belief-desire law).

Chapter 2 then turns to the way that contemporary **cognitive psychology** describes the causes of *action and action choice*. Cognitive psychology, in taking account of numerous predictable and systematic biases and cognitive illusions, has developed models of action and action choice in which the causal roles are not filled by beliefs and desires; these models have a number of features at odds with philosophical folk psychology.

Chapter 3 is concerned with theories of *interpersonal understanding* in contemporary **social psychology.** Addressing how individuals account for their own actions and the actions of others, the field of *attribution theory* employs models that do not involve the ascription of specific beliefs and desires, in direct contradiction of a key contention of philosophical folk psychology.

In both of these chapters, *unconscious biases of reasoning and cognitive illusions* feature prominently, as do the uses of **heuristics**. Proponents present the belief-desire law as both a **descriptive** account of human decision-making, interpersonal understanding and reason

giving and as a **normative standard** of rational action. When scientific psychology uncovers systematic divergences from this picture, their investigations of the processes and mechanisms by which these emerge shed light on the nature of the target phenomena.

Part Two of the thesis (Chapters 4-6) is concerned with the way that the terms "belief" and "desire" are used in everyday discourse. Although guided by some so-called "ordinary language"[30] approaches to philosophical reflection, the answer to the second guiding question emerges directly from examination of examples of the use of "belief" and "desire".

Chapter 4 offers an examination of how "belief" and "desire" feature in the construction of **narratives**. Along the way I discuss **narrative psychology** and **narrative psychotherapy** in order to draw attention to the powerful role that narratives play in the construction of deeply held convictions – even when these convictions have negative behavioural effects. I suggest that we might judge the *plausibility* of a relation between an individual's specific beliefs and desires and their subsequent actions on the basis of how closely a narrative that features such terms resembles culturally established archetypes – a possible source of the persuasiveness of the belief-desire picture of action and of its metaphysical assumptions. The suggestion that the propensity to ascribe beliefs and desires in action explanations, and the concomitant presumed relationship between propositional attitudes and action is a **cultural artefact** is supported by historical and anthropological evidence. Although the telling of stories seems to be a human universal, narratives without cognates of "belief" and "desire" in the functional roles defined by the belief-desire law are to be found in the ancient world and in some present-day cultures. In short, narratives involving belief and desire ascriptions are *not the only way that action is described in everyday discourse and when used in the construction of narratives, these terms do not usually pick out the causes of actions*.

Chapter 5 is concerned with the ways that people make **excuses** for actions and seek to **justify** those actions in cases where they are challenged or face potential censure. Using everyday examples and examples from the formal setting of legal procedure, the chapter develops, after Austin (1979a) in *A Plea for Excuses*, the excuse/justification distinction

[30] This term has gained a certain notoriety, both through over use and misunderstanding. To the extent that this approach is an "ordinary language" one, I do not mean to suggest that "everyday" uses of a term are *authoritative* as to its meaning. Instead, I suggest that close attention to the function and purpose with which a term is being deployed in a given circumstance can help shed light on potential misunderstanding about how it works and consequent "puzzling" aspects of its use.

and suggests that excuses, properly understood, do not constitute a claim about the *causes* of an action, even on those occasions that "belief" and "desire" feature. I argue that excuses that *do* advert to an individual's claimed beliefs and desires are neither offered as causal explanations of actions nor evaluated as to the causal efficacy of such states. I point out that safeguards have been put in place in the formal treatment of such excuses in the (English) legal setting to avoid the pitfalls of causal assumptions.

Chapter 6 looks at a further example of a type of discourse in which one of the essential terms of philosophical folk psychology, "belief", features without any commitment to its being a cause of action: **hedging.** The chapter begins by defining a **hedge** or a **hedging-phrase** and establishing that affixing "I believe that…" to a statement frequently serves this function. It is then proposed that although it is possible to read *hedges* featuring the phrase "I believe that…" as referring to the attitudes of the utterer, this can be accommodated *without* any commitment to the causal role of a particular belief, desire or pair of such attitudes within a network of causes. Once again, the state picked out by the phrase "I believe that…" is <u>not</u> a causal one.

The conclusion of the thesis, **Chapter 7**, examines some implications to be drawn from the foregoing. As well as taking the inferences from each chapter in turn, I suggest that once freed from the need to shape any competing ideas into the dogma of the belief-desire picture, the field is open for some new approaches to the questions of action choice and interpersonal understanding, as well as to the philosophical understanding of rationality more generally. Philosophers so motivated will find the elimination of the terms "belief" and "desire" from causal accounts in the philosophy of action, interpersonal understanding and reason-giving *key* to avoiding chasing solutions to illusory problems. Then we will have been successful in *bringing back these terms from their metaphysical to their everyday use*.

Part One:

“Belief” and “Desire”

in

Psychology

1 Chapter One:
Causal Explanation in Psychology

Abstract: Part of the therapeutic aim of this thesis is to undermine the conviction that actions are causally explained with reference to beliefs and desires. In pursuit of this aim, subsequent chapters (2 &3) will show how psychologists explain actions without recourse to propositional attitudes. For this to have force, it must first be established that psychology is concerned with genuine causal explanations: that is the objective of this chapter. I will contend that scientific psychology does not try to establish ***laws*** *or* ***law-like regularities*** *that govern action-behaviour; even if they did, describing a regularity would not constitute a causal explanation but, rather, an* ***explanandum*** *to be investigated. Psychological disciplines investigate the mechanisms that bring about regularities through a process of* ***functional analysis****. This term will be explained and described, together with examples of functional analysis at work.*

> … the essence of the mind being equally unknown to us with that of external bodies, it must be equally impossible to form any notion of its powers and qualities otherwise than from careful and exact experiments, and the observation of particular effects, which result from its different circumstances and situations.
>
> David Hume, *A Treatise on Human Nature*, (1740/1985: 44)

1.1 The Nature of Psychological Explanation

If the terms "belief" and "desire", as applied in philosophical folk psychology, successfully picked out the *genuine causes* of actions, or were used as causal terms when individuals seek to explain or predict their own or others' actions, we would expect psychologists to formulate **genuine causal explanations** in which these terms, or the belief-desire law, play a central role. In Chapters 2 and 3 of this thesis I will suggest that many cognitive and social psychologists offer explanations of the phenomena of action choice and interpersonal understanding, central to philosophical folk psychology, without recourse to these terms or the belief-desire law. This suggests that that belief and desire are not *essential* to

psychological explanations of action-choice or interpersonal understanding: other explanations are available. Before the significance of those observations can be understood, however, it is necessary to establish that scientific psychology is committed to genuine causal explanations of its target phenomena and to establish the form that these explanations take. The questions that this chapter deals with are:

1) What does it mean for scientific psychology to *explain*?
2) How are these explanations constructed?

In developing the answers to these questions, I will draw attention to a number of features of scientific psychology, including:

a) Cognitive and Social psychology seeks *genuine explanations*[31] for psychological phenomena and for the relationships between psychology and behaviour.
b) Such genuine explanations do not depend on the subsumption of psychological states and behaviour under *laws* or theories that invoke *law-like regularities*.
c) Scientific psychology proceeds by establishing *process models.*
d) These models are developed at ever more finely grained levels of explanation by means of *functional analysis*.

As a manifesto for scientific investigation in psychology, the quotation from David Hume at the opening of this chapter remains pertinent. Psychology is an empirical investigation into how humans process perceptions, memory, emotions, bodily feelings, language, and related elements of cognition both consciously and at the sub-personal level. Its inquiries begin by observing, for example, regularities between psychological states and consequent behaviour or consequent psychological states. The experimenter will control as *independent variables* the stimuli or other circumstantial conditions that evoke the particular psychological state. Alternatively, in the case of the observational (as opposed to experimental)[32] investigations, the researcher will observe and record variations in behaviour or self-reported psychological states (garnered by means of questionnaires, for example) under variable environmental conditions. These outputs are the *dependent variables* of the investigation.

[31] As distinct from merely predictive and data-fitting "*as-if*" explanations. See **2.2** below for an exposition of this distinction.

[32] For simplicity, this thesis is primarily concerned with experimental investigations.

This outline of the investigatory method, however, leaves open an account of the *explanatory strategy* employed by the psychologist as scientist. Such an account is the subject of this chapter.

The dominant investigatory paradigm in contemporary scientific psychology is an *information processing model* (Davey 2008: 259-62). Leaving aside the question of the extent to which the operation of the human mind resemble those of a digital computer (a frequently cited analogy)[33], this approach treats cognitive inputs (perceptions, memories etc.) as *information* over which processes of acquisition, storage and transformation are performed. These processes give rise to outputs which might be expressed as further cognitions, behaviours or even verbal reports. The goal of scientific psychology is to describe those operations in terms of detailed *process models*. An accurate process model would be proposed both to explain the observed regularities and to predict future outcomes under given cognitive conditions (although see below, regarding how explanation and prediction come apart). Thus, and importantly for the discussion to follow, the objective of psychological explanation goes beyond the mere recording of regularities and into describing *how* those regularities arise.

In the fundamental natural sciences, like physics and chemistry, identifying the *laws* – perhaps even the *laws of nature* that govern the circumstances under which a phenomenon occurs (as distinct from those under which it does not) might be sufficient to *explain* an individual occurrence of that phenomenon. Prediction, in the case of a law-governed scientific model, would be a matter of stating the governing laws, together with a description (real or hypothetical) of the prevailing conditions. Such laws might be described as *universally quantified material conditionals* (Oaksford and Chater 2010: 6). If psychology were to explain in much the same way as the fundamental natural sciences, then the objective of scientific psychology would be to discover the *right set of laws* and to define which psychological phenomena they govern. In the middle years of the 20th century, this view of psychology had its adherents. For example:

[33] In the sense that I use it here, "information processing" is neutral with regard to this question and well-rehearsed philosophical discussions regarding the nature of "representations".

> If psychology is to become a natural science, it will have to formulate and use some set of causal laws that are at least consistent with the causal laws of the other natural sciences.
>
> (H. M. Johnson 1939)

Although Johnson went on to argue that the laws of psychology need not be reduced to physical (or biochemical) laws, the relation between laws and observed regularities and the assumption that a law-like regularity must be put forward in order to facilitate the prediction of psychological phenomena has been present in much of psychology's history. It might be argued that the formulation of theories deploying law-like regularities between *stimulus* and *response* drove the turn to behaviourism that dominated (especially) North American psychology during the middle years of the 20th century until the so-called "cognitive revolution" of the 1960s. Consequently, it has also been a persistent picture in the philosophy of psychology.

This is the picture that Cummins (1983, 2006) challenges. He suggests that regarding psychological explanations as depending on a *"covering law model"* – of the kind proposed by Hempel and Oppenheim (1948) – is to mistake the *explanandum* of a psychological observation for the *explanans*. Observed regularities between stimulus and behaviour, between psychological states and behaviour or between psychological states and further elicited psychological states are *effects* that remain to be explained. Indeed, offering an "explanation" of a series of regular occurrences, or an individual occurrence that is consistent with that regularity, by appealing to a "law" is to do nothing more than to restate the regularity.

For example, imagine an experiment under which a psychologist asked subjects to identify, by species name, a sequence of images of animals. For the duration of the task, which requires them to speak the name of the species aloud as soon as they have identified it, the subjects wear a set of audio headphones. In the first run of the experiment, the subjects perform the task while "white noise"[34] is played through their headphones. In the second run, they are asked to perform a similar task while listening to the voice of an actor narrating

[34] A sound similar to that of radio "static" in which all audible frequencies are played simultaneously at similar amplitude.

a story. A measurement of their *response time* (the delay between the image being shown on the screen and their correctly identifying the species) is recorded.

Suppose that analysis of the response-time data suggests that, on average, subjects take longer to retrieve species names while listening to the story than while the white noise is being played[35]. The experimenter might be tempted to record his findings as:

> L(i): Exposure to sound of the human speaking voice tends to increase memory-retrieval response time compared with exposure to non-verbal sounds of similar amplitude[36].

Subsequent replications of the experiment produce consistent results. The "law" described at L(i) seems to hold. I would contend, however, that we would be reluctant to say that L(i) *explains* the observed regularity. In essence, it merely restates it, as Cummins and Bechtel and Wright (2009) would insist. Claiming that, statistically, people tend to retrieve knowledge from memory more slowly when they are listening to speech sounds than when they are listening to something else would be an interesting effect (if true) and would be worthy of further investigation. However, it *explains* nothing. As an effect it would be a *new explanandum, a phenomenon in need of explanation*.

A psychological regularity like that at L(i) stands in need of more finely grained examination in order to uncover its underlying structure and the *mechanisms* that bring it about. The experimenter might begin by asking whether the effect occurs only when the subject is exposed to the spoken word as coherent narratives. Does it persist if the sound is a repeated sentence, a repeated word or even a sequence of non-verbal speech sounds? And in the latter case, does the kind of sound – consonant or vowel, plosive or sibilant – make any difference? What about the sounds of speech in languages that the subject does not understand? The experimenter would also want to vary the *dependent variable* – that is, the information than the subject is asked to recall, not only by asking them to retrieve information other than species names of animals but other data instead – types of vehicle, or tool, or to match names to famous faces but also to describe *textures,* the cue for which is acquired through the haptic modality, to investigate whether the effect is limited to an effect of auditory stimuli on visually cued memory retrieval. These are just some examples.

[35] I have not based this example on any real experiment so this "result" is purely speculative

[36] Loudness.

The objective of redesigning and refining these experiments is to uncover the structure of the explanandum and thereby to get ever closer to an *explanans*. This would not be another law-like regularity but an analysis of the functional relationships underpinning L(i). This brief description, which, in reality, might comprise years of investigation and countless experimental findings, is a crude characterisation of the process of *functional analysis*. Cummins (2006: 96) describes this process thus:

> Functional analysis consists in analysing a disposition into a number of less problematic dispositions such that programmed manifestation of these analysing dispositions amounts to a manifestation of the analysed disposition.

The "analysed disposition" described here is the regularity – such as that at L(i) – to be explained. "Less problematic dispositions" are those regularities which occur at ever more simple, subpersonal levels of psychological explanation. By "programmed manifestation" Cummins is pointing out that a satisfactory analysis does not merely break the overall disposition into a dissociated set of smaller functions but also seeks to describe how these functions interact, their systematic relationships by means of which their individual effects and effects on each other bring about the "analysed disposition".

Bermúdez (2005: 63-69) suggests that the ultimate goal of subpersonal functional analysis is to arrive at a set of functions which are one-to-one reducible to the underlying physical (neurological, neurochemical and neuro-electrical) events. It is a conceptual matter whether this endpoint is, in principle, achievable. Less controversial is the contention that functional analysis allows the scientific psychologist to go beyond "what happens when…" descriptions and to describe how the effects that are to be explained come about. An example from a more tangible, mechanical domain might make this idea clearer.

If one were to ask how a *pump* works – or even how a *particular* pump works – the questioner would be unlikely to accept the statement "fluid enters at one point and leaves at another under greater pressure" as an *explanation*.[37] The question was not "what does the pump do" (which could be rendered as "what is its gross function") but *"how does it work"* or, to put the same question another way "what are its components and how do these combine with each other in order to perform the work of the whole as a *pump*?"

[37] Note, however, that this is exactly the form of "explanation" offered by L(i).

We might want to know how the pump is powered, what it is made of, its capacities and other dimensions. To go further, we would need to disassemble it into its components. We could then begin to identify how its operation brings about its overall effect (analysed disposition). If we want to know, in the case of an individual pump, how it functions qua pump, we are going to have to take a screwdriver to it!

Once we have dismantled enough pumps we might begin to notice some regularities at the level of our decomposition. Some pumps have rotary impellers; some have reciprocating pistons. Some are powered by an electric motor, others by internal combustion engines or even steam power. To answer our "how does it work" question, however, we might not need to know, necessarily, how the reciprocation of pistons or the rotation of an impeller imparts pressure to a stream of fluid. Although we could analyse the component operation to the level of fluid dynamics, explanations of how a particular pump works could become satisfactory before we reach this point. Describing the input and output valves, the way that the impeller or pistons are driven - including the method of transferring the drive to these components and regulating their speed – is likely to be a satisfactory answer to the question of *how the pump works* – or how this specific type of pump works – for the mechanic, if not for the theoretical physicist.

Thus "how does it work" questions can be satisfied at different levels of explanation. The mechanic might ask the question because they want to be able to repair a faulty pump. Reduction to the underlying principles of physics is not necessary to this endeavour. Neither would the gross "disposition" of the pump give a clue of how to begin a repair, unless we want to risk a situation in which a mechanic orders a new set of piston seals to repair a faulty impeller-driven pump.

In the case of psychology, the acceptable level of explanation is likely to be quite different for the clinical psychologist (who needs to know how things should work in order to effect a "repair"), the investigative, theoretical psychologist (who wants to understand the mechanisms that give rise to regularly observed dispositions), the neuroscientist (who wants to know how these operations are instantiated in neurological structures) and the biochemist (who wants to understand the reactions and chemical kinetics that go on below the level of the individual nerve cell). None of these, however, will be satisfied with the kind of regularity described at L(i) any more than a pump mechanic will be satisfied with "low

pressure fluid in: higher pressure fluid out". All would demand a **functional analysis** to answer the question "how does it work?"

1.2 "*As If*" versus Genuine Explanations and Explanation versus Prediction

Gross dispositional descriptions might closely fit the data collected by many observations. All of the (working) pumps you examine might exhibit the "low pressure in, higher pressure out" regularity to different degrees. Plotting these findings on a graph would clearly show that there is a correlation between the presence of a pump and a pressure differential on either side. Likewise, plotting the results from the hypothetical psychological experiment described above would show a statistically significant correlation between the type of auditory stimulus and subjects' response times. There might be significant outliers, but these could be disregarded without doing damage to the key finding, as described at L(i).

Psychological studies might generate any number of these statistically significant correlations but, as we have seen, these are not to be safely regarded as explanations. Any such finding could be described using the phrase *"as if"*. The statistical outcome is *"as if" people have more difficulty remembering the names of animals when listening to speech*. However, a good deal more functional analysis is required if we are to justify this "as if" hypothesis by means of sequences of causes and effects. When the mechanic functionally analyses a pump they are seeking a series of causal relations which, working together ("programmed" in Cummins' word) *cause* the fluid to leave the device at a higher pressure than it had on entering it. When the psychologist seeks to break an observed phenomenon down into its component functions by means of functional analysis they are looking for the *causes* of the observed data. Here I am using "causes" in a perfectly ordinary, everyday sense. One need not hold a view on the metaphysics of causation to understand the straightforward sense in which psychologists seek to understand causes. Their goal is to describe the mechanisms through which phenomena arise, rather than restating or repackaging those phenomena or by constructing hypothetical "as if" statements which, *although they fit past data, remain to be explained in terms of the causal history that ensures their conformance*.

Frequently we find "explanation" and "prediction" used together as if they are simply two sides of the same process, differentiated only by their temporal relationship to the observed phenomenon: explanations are offered after the event, predictions are temporally prior. A single law-like relationship, like that at L(i) is frequently presumed to be capable of serving

as either explanation or prediction. The "law" *predicts that future findings will be in accordance with it* (in this hypothetical case, "people will respond to visual stimuli more slowly when listening to speech") as well as *explaining why an individual result conforms to the law* by means of a statement like "this is explained by the established fact that people have more trouble remembering the names of animals when listening to speech." It is, after all, a **law**.

This picture is, however, scientifically inaccurate. A statement of a law-like regularity predicts almost nothing about what will happen if the parameters of a given experimental investigation are altered – for example, in ways similar to those described above. Scientists might want to make testable predictions about what will happen in these circumstances; doing so, however, will entail the proposal of an additional hypothesis. When the results support the hypothesis or radically falsify it, the scientist has *additional information* with which to continue testing their hypothesis or is aware that a new hypothesis is required. The essential asymmetry between explanations and predictions in science lies in the fact that *explanations* – whether they support or reject a given hypothesis – always come equipped with information that is not available at the time the prediction, or hypothesis formation, is made. This information is usually in the form of new experimental or observational data.

The psychological literature is replete with examples of the use of functional analysis in the development of genuine explanations from which I draw two illustrations. The first is David Marr's analysis of the computational functions of the human visual system (Marr 2010). The second is the historical development of the dominant contemporary account of **semantic memory**; the **spreading activation** model.

1.3 Functional Analysis in Practice: Marr's *Vision*

David Marr's 1982 work *Vision* is a seminal example of functional analysis. He begins by setting out the territory for his investigation and something of the conceptual landscape by offering a definition of the target phenomenon:

> …vision is the process of discovering *from images* what is present in the world and where it is.
>
> Vision is therefore, first and foremost, an information-processing task,
>
> (Marr 2010: 3 emphasis added)

He is at pains to point out, however, that understanding the *process* is only part of the task of understanding the means by which visual information is made available to cognition. Equally important is understanding how visual information is *represented* in the cognitive apparatus of the "seeing" creature.

> The study of vision must therefore include not only the study of how to extract from images the various aspects of the world that are useful to us, but also an enquiry into the nature of the internal representations by which we capture this information and thus make it available as a basis for decisions about our thoughts and actions.[38]

(Ibid.)

His "quite general" definition of a representation, of what it means for image information to be captured and stored and for that information to be available for and subjected to transformations, is another example of his computational model.

> A *representation* is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this. And I shall call the result of using a representation to describe a given entity a *description* of the entity in that representation.

(Ibid. 20)

Marr argues against the impression that this "computational" (information processing) focus is tantamount to reducing human experience to the architecture of a traditional computer. Computation, as a process, is not the sole domain of the human-made artefact that we call a computer. Explaining that human experience can be described in terms of the sequence of computational processes that give rise to it is not the same as "reducing" experience to the operations of a "mere computer", Marr argues that:

> To understand a computer, one has to study that computer. To understand an information-processing task, one has to study that information-processing task.

[38] Some of this language is philosophically contentious; not least the phrase "internal representations", which raises the issues both of the internal/external division and of the nature of mental representation. Such concerns can be set aside, however, as they have no bearing on the analysis of Marr's method.

> To understand fully a particular machine carrying out a particular information-processing task, one has to do both things. Neither will suffice alone.
>
> (Marr 2010: 5)

This observation informs much of Marr's approach to understanding vision. It is not sufficient to posit a representational theory or to describe a computational process unless these can be reconciled with what is known about the physical, neurological apparatus in which they are instantiated. At the same time, no explanatory description of any of these processes – representational, computational, neurological – can be regarded as safe unless it is compatible with the ordinary experience of visual phenomena.

Marr breaks his analysis of visual processing into three levels: the *computational*, under which the tasks to be performed by the visual system are themselves broken down into a series of mathematical functions that must be performed; the *algorithmic* level, which asks how the computational functions are implemented – what, for example, are the representations involved and what are the algorithms for the transformation of acquired images into usable information and finally the *implementational* level which seeks to describe how computational and algorithmic levels are instantiated in the material of the human visual system – retinas, nerves and brains, for example (Bermúdez 2010: 48).

As an example of Marr's analysis of the computational task he begins by separating out the four factors "responsible for the intensity values in an image" (Marr 2010: 41) which he identifies as the object's (and the captured image's) *geometry*, *reflectances* (of the visible surfaces), *illumination* and (the subject's) *viewpoint*. This allows him to postulate how the processes of the visual system are able to differentiate between changes in the values of each of these factors. Within each of these the most fundamental change in intensity to be detected by the visual system is a "zero-crossing" (Ibid. 54) where the value of the intensity of a stimulus[39] passes from positive to negative or vice versa. This provides "a natural way to move from an analogue or continuous representation … to a discrete, symbolic representation." (Ibid. 67). It also allows Marr to give a computational account for one of the fundamentals of visual acquisition: *edge detection.* Marr avoids mention of "edges" for as long as possible because of the word's physical connotations beyond the consideration of image construction (Ibid. 68) Objects have "edges"; in the case of images there are only

[39] In common with any mathematical function.

variations in intensity and "zero-crossings" between adjacent but discernible changes in intensity. An individual image can be broken down into a series of zero-crossings, as shown in this example (Ibid. 69):

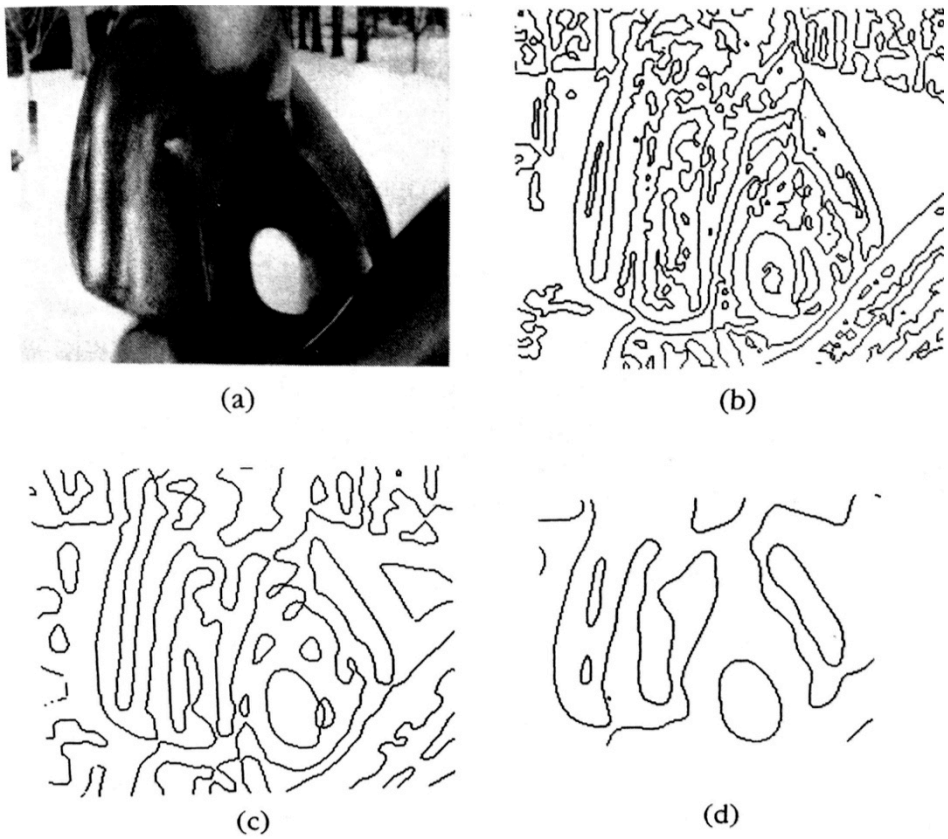

Fig: 1.1: Marr's demonstration of how a complex shaded image can be rendered as a series of "zero crossings".

Zero-crossings are the first of a series of "primitive" computational tasks that Marr identifies, which taken together help the visual system to build up a *primal sketch* of the scene (Marr 2010: 37). Primal sketch information is processed with information about local surface orientation, distance from viewer, discontinuities in depth and discontinuities in surface orientation to produce a $2\frac{1}{2}$*-D sketch* which Marr claims "makes explicit the orientation and rough depth of the visible surfaces, and contours of discontinuities in these quantities in a viewer-centered coordinate frame" (Ibid.).

Throughout, Marr is guided by his central concern with computational theory. This is essential to his functional understanding of the visual system because "the nature of the computations that underlie perception depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented" (Marr 2010: 27).

Not that this concern with the mathematically expressed fundamentals of vision leave him with an esoteric investigation far removed from visual experience. Among the "computational problems" that he must account for are the findings of many contemporaneous experiments on the visual systems of real people. Likewise, he is fascinated by the effects of optical illusions – especially illusions of shading, such as this:

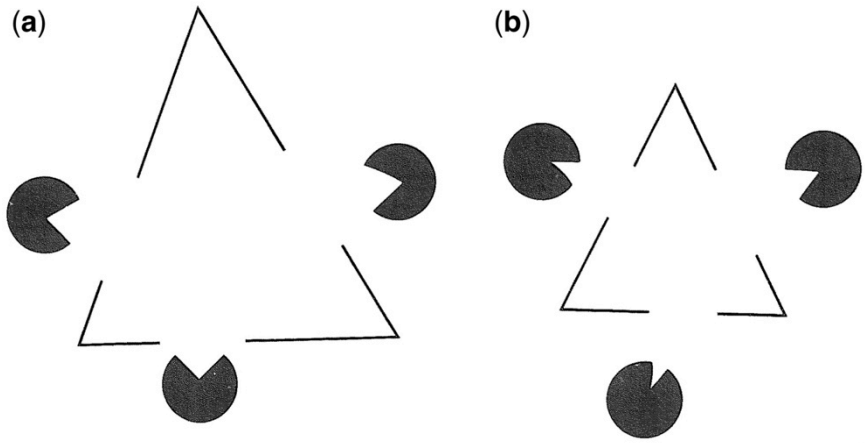

Fig. 1.2 Marr's illustration of subjective contours (Marr 2010: 51)

Throughout the work, Marr offers a description of the experimental and observational investigations that inform his analysis. Each of these takes an individual function and breaks it into its computational processes.

Marr offers a detailed defence of his *approach* in Chapter 7 of the book, including an imaginary "conversation" with an interlocutor whose questions are drawn from his experience of lecturing on his ideas. Among them, he is asked whether his "different levels" of explanation are really independent[40]. He responds:

> Not really, though the computational theory of a process is rather independent of the algorithm or implementation levels, since it is determined solely by the information-processing task to be solved. The algorithm depends heavily on the computational theory, of course, but it also depends on the characteristics of the hardware in which it is to be implemented. For instance, biological hardware might support parallel algorithms more readily than serial ones, whereas the reverse is probably true of today's digital electronic technology.

[40] One might substitute "autonomous".

(Marr 2010: 337)

More than thirty years after Marr's (posthumous) publication, many of his conclusions have been superseded and a number of the *computational problems* that he identified are still to be satisfactorily described (Gazzaniga et al. 2009: 217-25). Nonetheless, Marr's approach, of breaking the description of the visual system down into a series of computational functions, treating each individually and postulating ways that they causally interact, informed by experimental results and the findings of cognitive neuroscience, remains the paradigmatic way that the visual system is investigated.

> Marr's computational approach to vision, which distinguishes the overall task of vision from the algorithms for its solution and from its neurophysiological implementation, exemplifies one predominant contemporary approach to perception in cognitive science.
>
> (O' Callaghan 2012: 83)

> Information is represented on multiple scales. Although early visual input can specify simple features, object perception involves intermediate stages of representation in which features are assembled into parts. Objects are not determined by their parts; they are defined by relations between parts. An arrow and the letter Y contain the same parts but differ in their arrangement.
>
> (Gazzaniga et al. 2009: 225)

Although the precise computational processes instantiating this assembly and differentiation remain to be described, the method by which they will be uncovered is likely to follow Marr's approach. This method is central not only to vision research but to most other areas of psychological investigation.

1.4 Spreading Activation in Semantic Memory

Explaining the operation of the *human memory* in *information-processing* terms requires a theory that encompasses all of the familiar experiences of memory, including forgetting, or the felling of having some name, word or fact "on the tip of the tongue". Additionally, psychologists have encountered less familiar effects, such as **priming**: i.e., that it easier and faster for a subject to recall facts when some connected information has been made salient

prior to the test[41] (Baddeley et al. 2009: 81-82; Neely 1977). The priming effect suggests a way into the investigation of memory: by varying the nature and presentation of primes researchers have a quantifiable tool with which to investigate the computational processes involved in retrieval. Priming effects and many other features are consistent with the idea that semantic memory is associative and that retrieval from memory operates across a **spreading activation network**.

Theorists describe the long-term[42] human memory system as comprising two parts: **episodic memory** (Tulving 2002), of which autobiographical memory is a part,[43] is concerned with storing and retrieving information about **events** that the subject experiences. This will include memories of the elements of **subjective experience** – what the individual saw, smelled, felt and even the emotions that prevailed at the recalled moment. **Semantic memory** does not encode this subjective experience. It is concerned with the storage and retrieval of **knowledge** about the world; knowledge of facts, categories, meanings, signs and symbols. It is also where rules for the use of words and symbols and learned knowledge of problem-solving techniques are stored. As Baddeley et al. (2009: 114) describe it, semantic memory, as the name suggests, also facilitates the linguistic ability to deploy words in meaning-appropriate ways.

> [Semantic memory] is a mental thesaurus, organised knowledge a person possesses about words and other verbal symbols, their meanings and referents, about relations among them, and about rules, formulas and algorithms for the manipulation of these symbols, concepts and relations.
>
> (Tulving 1972: 386)

Imagine hearing your grandfather telling you stories about his life as a young man; the work he did, the places he visited, the people he knew and even details such as the way that he dressed. For your grandfather, these reminiscences are drawn from *episodic memory* – as, for you, will be the memory of your grandfather telling you these stories. The information that he related would, however, become part of your *semantic memory*. It becomes, for you,

[41] Or more difficult in the case of "negative priming".

[42] As distinct from *short-term* or *working* memory.

[43] Some theorists prefer to regard autobiographical memory either as a separate system or as a distinct subsystem (Baddeley, 2009).

a series of facts the retrieval of which "lacks this sense of conscious recollection of the past" (Tulving 1972: 387).

Qualitative differences would not entail that the two categories of memory depend on distinct systems. Empirical evidence, however, suggests that they do. For example, brain-damaged people with *retrograde amnesia* – loss of memories from before the trauma frequently have deficits of episodic memory: they cannot remember specific events from their past. At the same, their semantic memory remains largely intact (Spiers et al. 2001; Wood et al. 2014).[44] They may not remember the sights, sounds, smells and emotions of their wedding day, for example, but they still understand what the word "wedding" means and are likely to recall the date and place at which the ceremony took place. To take our grandfather example, such amnesia would mean being unable to remember the occasion on which your grandfather told you stories from his youth while still remembering *what* he told you. That the deficits resulting from damage to particular areas of the brain are differentiated in this way strongly suggests that different neural systems are involved in the storage and/or retrieval of semantic and episodic memories.

Note, however, that information drawn either from episodic or semantic memory might be labelled "beliefs". Your grandfather might "believe" that he once wore a top hat (from his episodic memory: you, having been told this, would "believe" that your grandfather once wore a top hat – from your semantic memory. That psychology indicates that different systems are involved in the coding, storage and retrieval of "beliefs" of different kinds suggests that the defining "belief" as a unitary functional state is unwarranted.

The dominant theory of the psychology of semantic memory is the **spreading activation model**. The fundamental feature of this model is that the encoded elements that make up our concepts – nodes in semantic memory (A. M. Collins and Loftus 1975) – are linked to associated elements and concepts and the strengths or weights of the associative links determine the ease, speed and accuracy with which information is retrieved. Thus measuring **retrieval times** of associated information when particular nodes are activated by a **perceptual stimulus** can help the researcher to map the network of semantic nodes that make up an individual concept (McNamara 1992).

[44] According to the same authors, the reverse effect, loss of semantic memories with the retention of episodic recollections is less common. This has been taken as evidence of the robustness of semantic memory and supports a distributed rather than localised model, such as spreading activation.

An outline of the history of research into semantic memory and of how the spreading activation model has been developed is highly illustrative of how scientific psychology makes use of functional analysis in explaining phenomena.

One of the earliest investigations into the nature of semantic memory, and one of the first systematic models of the capacity (Baddeley et al. 2009: 116) was the *hierarchical network* model suggested by A. M. Collins and Quillian (1969). This paper investigated how people determine the truth of statements such as "a canary can fly". Collins and Quillian suggested two possibilities: the fact that a particular species of bird *can* fly might be stored in semantic memory along with the name of that species and that information would be repeated for every instance of a species of bird that can fly. This model, however, would require a great deal of duplication of information – as many instances of "can fly" as there were remembered species of flying bird – and so lacked *cognitive economy*, which the researchers assumed was likely to be a real feature of human semantic memory. Collins and Quillian's preferred model proposed that information true of "birds" in general would be stored alongside that "higher-level" category, and derived from that to its members. Diagrammatically, the hierarchical model looks like this:

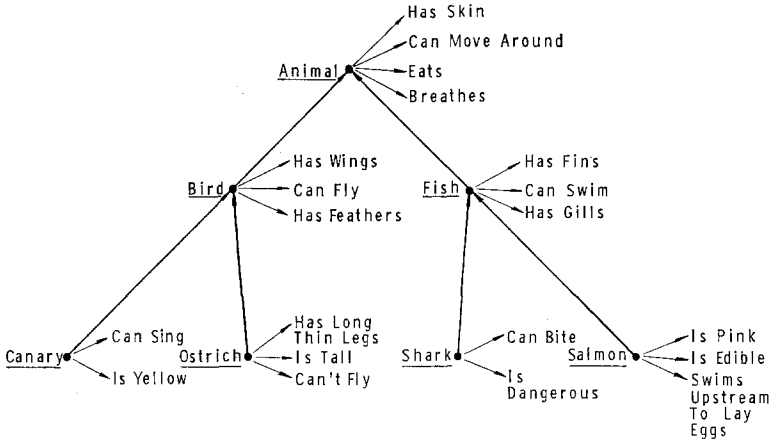

Fig. 1.3: Diagrammatic representation of hierarchical model of semantic memory, reproduced from Collins and Quillian (1969).

Information specific to the individual member of a group is stored alongside the name of that species. Thus "is yellow" is stored at the same level in the hierarchy as "canary". Information that pertains to the category as a whole is stored at the category level: "has wings" belongs with "bird" because this is true of birds in general and not just of canaries.

To put this model to the test, Collins and Quillian devised a series of "true or false" questions which were put to a number of subjects whose responses were timed. The hypothesis was that if the information encoded in the target statement required retrieval from different levels of the hierarchy, response times would be longer than if the information came from the same level. This would mean that subjects would take longer to identify "a canary has skin" as true, than they would to assent to "a canary is yellow".

Collins and Quillian's results were much in line with this prediction.

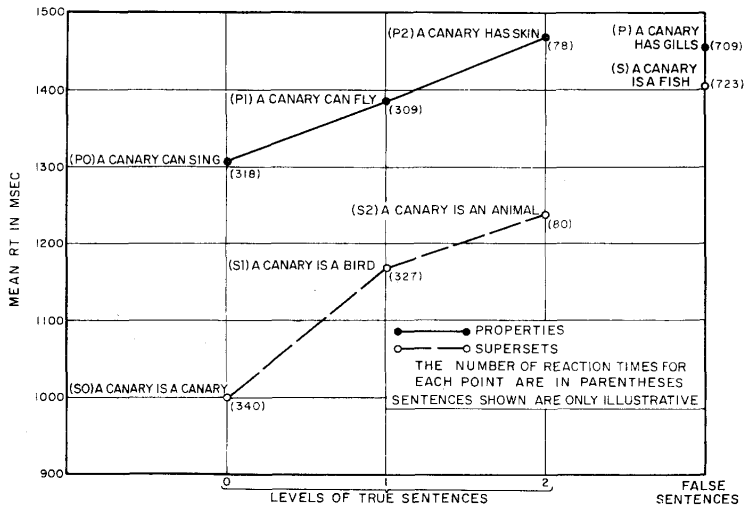

Fig 1.4: Plot of response times from Collins and Quillian (1969)

These results show that subjects, on average, took longer to identify "a canary has skin" as true than they did "a canary can fly" which in turn took them longer than "a canary can sing".

Although these results were suggestive that the hierarchical model was correct, critics pointed out that there was another possible explanation for these results. Associated information, such as "a canary can sing" (or even the phrase "singing canary") are familiar tropes. Perhaps, the sceptical argument suggested, closely allied ideas are retrieved more quickly simply because they are so familiar. This was investigated by Conrad (1972), who found, according to Baddeley et al. (2009: 118) that "when familiarity was controlled, hierarchical distance between the subject and the property had little effect on verification time."

The hierarchical model also fails to account for another feature of semantic memory. Experiments on response times like those carried out by Collins and Quillian found that

statements such as "a canary is a bird" tended to be more rapidly identified as true than less typical examples, such as "a penguin is a bird" (Ibid.). The "hierarchical distance" is the same in each of these cases. A complete model of semantic memory would have to account for these "typicality" effects.

In direct response to this, Rosch and Mervis (1975) developed a model of categorisation based on *prototypes* that directly referenced the notion of *family resemblance* suggested by Wittgenstein (2009 §67). On this model, the association of an example with a given category is dependent on how closely that example resembles a prototypical member of that category. The effect that this model suggests has been shown to be robust through many replications and variations on Rosch and Mervis' original investigation. However, it would have remained an *effect, a regularity yet to be explained* had not A. M. Collins and Loftus (1975) suggested a mechanism by which semantic memory acquires, stores and retrieves conceptual, categorisation and factual information. This was a much more flexible model than hierarchy alone and accounted for typicality effects, being based on semantic, rather than hierarchical, distance. This was the **spreading activation model of semantic**

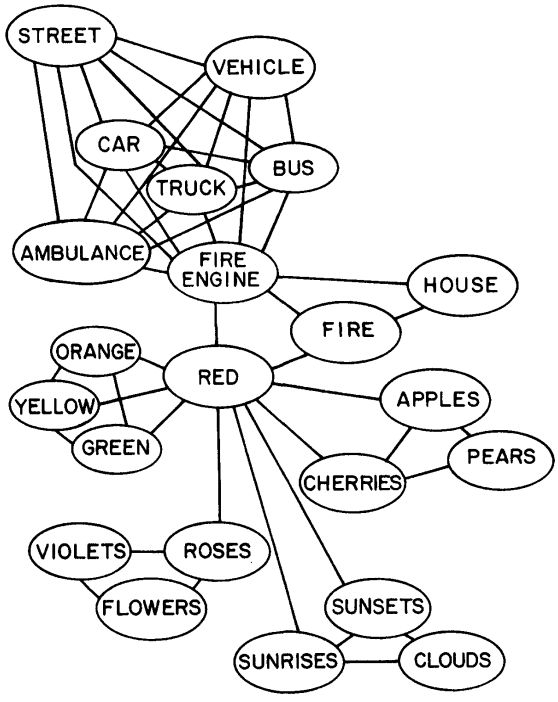

Fig. 1.5: Diagrammatic representation of concept relationships under a spreading activation model. From Collins and Loftus (1975).

memory.

In this diagram, the length of the lines between *nodes* – individual concepts in memory – represents the strength of association. The concept "fire engine" at the centre of the matrix is closely associated with the concepts "ambulance" and "fire" although less strongly associated with "vehicle". This would mean that a stimulus that evokes "fire" is likely to activate "fire engine" much more readily and rapidly than would an evocation of "vehicle". Activation of items in semantic memory spreads from node to node,[45] dependent on the strength of association until the required concept, fact or idea is retrieved.

A. M. Collins and Loftus (1975) describe a series of experimental findings that map on to their model quite closely. For example, Loftus (1973) had investigated category-instance evocation. Four kinds of pairing between an instance and a category were tested by asking the subject whether the example was a member of that category (e.g. "is an oak a tree?"):

> … (a) pairs where both the category and instance evoked the other with high frequency (e.g., "tree-oak); (b) pairs where the category evoked the instance with high frequency, but the instance evoked the category with low frequency (e.g., "seafood-shrimp"); (c) pairs where the category evoked the instance with low frequency, but the instance evoked the category with high frequency (e.g., "insect-butterfly") and (d) pairs where both the category and the instance evoked the other with low frequency (e.g., cloth-orlon).
>
> (A. M. Collins and Loftus 1975)

The presentation of each pairing was varied as to whether the category or the instance was presented first. As predicted by the spreading activation model the findings were that:

> …subjects are fast when the category is presented first, if the category evokes the instance with high frequency, and subjects are fast when the instance is presented first, if the instance evokes the category with high frequency.
>
> (Ibid.)

It is thus semantic distance (frequency of association, represented by the shortest route along the meaning-pathways) that determines how quickly a pairing is recognised rather than

[45] This is not the same as a connectionist model of brain architecture: a "node" in these models does not equate with a single brain location. Spreading activation is a model of conceptual memory, not of brain function. (McRae & Jones, 2014: 207-208)

lexical association alone. The findings can, they suggest, be explained by spreading-activation:

> …activation spreads along some number of pathways, because the subject has activated the lexical network in addition to the semantic network. Hence, in the present explanation, the subject's control is reduced to diffusely activating whole networks rather than specific pathways (in addition to the specific nodes activated by the stimuli in the experiment).
>
> (Ibid.)

Over the ensuing decades, researchers have honed and refined this image of the spreading activation model by isolating individual features and testing them experimentally to gain an overall picture of the processing of semantic information. In his review of much of this research, Elman (2009) suggests that one result has been a shift of focus from the *rules* governing word use to the "semantically rich" content of individual words themselves – words that gain much of this richness from the web of associations in which they are held, rather than from a lexical "look-up" process. This implies that words are not only…

> …flesh that gives life to grammatical structures, but [also] bones that are themselves grammatical[ly] rich entities. This sea change has accompanied the rise of usage-based theories of language which emphasize the context-sensitivity of word use.
>
> (Elman 2009)

However, one feature of our semantic memory that is predicted by the spreading-activation model is that we are prone to semantic errors, particularly when incorrect information activates closely associated information (Erickson and Mattson 1981; Hannon 2014).

1.5 Developments of Associative Theories

The associative implications of the spreading activation model have led other researchers in new directions. For example, Morewedge and Kahneman (2010) develop a new take on the dual process model of judgement and decision making (see Chapter 2, section 2.7). They identify "intuitive judgements" with the operations of "System 1" (their term) – the fast, largely effortless and automatic processes are both essential to (and frequently sufficient for) many of our day-to-day decisions and yet prone to "systematic errors" or biases.

For example, take the illusion suggested by Erickson and Mattson (1981). If we ask some subjects the following question:

> How many animals of each kind did Moses take into the Ark?

According to Erickson and Mattson (1981), and in subsequent studies reviewed by Park and Reder (2004), most respondents answer "two". This solution is recovered from memory without conscious consideration: without, in Morewedge and Kahneman's terms, invoking the more reflective "System 2". A moment's thought reveals that it was Noah, and not Moses, who, in the Bible story, took animals on to an ark.

This is explained by an associative account of automatic processing, and specifically the feature that Morewedge and Kahneman identify as "associative coherence". The associations of the names "Moses" and "Noah" are sufficiently similar – both are old-testament patriarchs whose story would have been learned in childhood by anyone raised within the dominant Judeo-Christian tradition of the West – that many will not notice the substitution of one patriarch for another. To confirm this, if one asks "How many animals did *Adam* take on to the ark?" almost nobody will fail to notice the substitution (Erickson and Mattson 1981). Although "Adam" is the name of the *original* Old Testament patriarch, the associations of that name are sufficiently distinct from those of "Moses" and "Noah" that respondents will "engage System 2" (Morewedge and Kahneman 2010).

Another feature identified by Morewedge and Kahneman is **attribute substitution**. When asked to identify pairs of rhyming words, we might not be surprised that subjects would pick out "VOTE-GOAT" more slowly than "VOTE-NOTE" in a written test (Seidenberg and Tanenhaus 1979). However, that a similar delay pertains when they *hear* pairs of words being spoken, indicates that the association of a word with its spelling is "…evoked automatically, although it is disruptive." Glucksberg et al. (1982) showed that when subjects were asked to state *whether a sentence was literally true*, they responded more slowly when there was a chance that the example was *metaphorically true* – for example "some roads are snakes" or "some jobs are jails" were identified as false more slowly than were other cases for which a metaphorical reading was not available.

A third feature of intuitive judgements invoking associations is described by Morewedge and Kahneman as **processing fluency**. In the example that they draw from Schwarz et al. (1991), participants were asked to provide either six or twelve examples of occasions from

their own lives in which they had acted *assertively*. When subsequently asked to judge themselves on a scale of assertiveness, the group who were asked for twelve examples consistently rated themselves *less assertive* than those who had been asked only for six. It is suggested that the difficulty associated with coming up with those last few examples of their own assertive behaviour tended to depress their impression of themselves as assertive individuals.

These observations have two significant implications. Firstly, it shows how a productive model of psychological processing – in this case associative activation – can continue to be informative and to suggest fruitful avenues of enquiry in areas beyond its original domain; this is surely the mark of a genuine explanation, rather than a merely *"as if"* predictions. Secondly, these examples bring out the importance of explaining the process mechanisms that underlie specific systematic errors. This is an important guide in scientific psychology's quest for explanations and is a theme to which we will return over the course of the following two chapters.

In common with Marr's approach, the spreading activation and associative models are informed by a quest to describe *processes*. Systematic biases (like the Moses illusion) are helpful in this respect because they help researchers to map out the computational steps that the subject goes through in retrieving and manipulating information – what are, for example, the circumstances under which substitutions are unnoticed? *How is retrieval time affected by different kinds of conceptual distance*?

1.6 The Scientific Status of Psychology

The adherent of philosophical folk psychology might raise the objection that one should not treat psychology – a **social science** – as if it had the same level of commitment to genuine, causal explanations as the **natural sciences**. This doubt lies on a continuum between two extremes: at one lies the view that psychology is <u>not</u> amenable to the same methodology as natural science and, at the other pole, the normative ethical judgement that psychologists ought not to carry out experimental investigations on human subjects.

This is significant for the purposes of the present thesis because, in challenging the scientific status of psychology, the adherent of philosophical folk psychology could claim that the models developed in the former have no more explanatory power than, for example, the belief-desire law. In this section, I will show why this challenge fails.

Concerns over its scientific status have haunted psychology ever since its establishment as a discipline. Chung and Hyland (2012: p76) point out that "Almost every introductory textbook on psychology starts with the statement that psychology is a science. The fact that this statement is made suggests that it is sometimes questioned." In one psychology textbook for undergraduates embarking on psychology courses (Davey 2008) an entire chapter (Ch3, pp20-31) is devoted to the debate. Those in favour of "psychology as science" point out the methodological rigour and investigative objectivity with which contemporary psychologists pursue their explanations (Bell 2002: 75-100). In response, sceptics who challenge the scientific status of the discipline claim that many of the theories put forward by psychology are **not falsifiable** (ibid. 82-3), in the terms defined by Popper (1992). As per Popper's charge against Freudian psychoanalysis, they argue that many of the theoretical posits of psychology are shaped to fit the data.[46] As such they could account for any new observations; they make no "bold predictions" that might be shown to be wrong[47].

If we accept the view that psychology seeks to uncover the *mechanisms underlying observed regularities* this objection is allayed. For the purposes of this thesis, I will take seriously the claim that psychology, by means of functional analysis, seeks to understand person–level psychological phenomena by uncovering the systematic causal relationships between the **subpersonal functions** that make them up – as proposed by Cummins (1980, 1983, 2006) and Bechtel and Wright (2009). Psychology, understood this way, is committed to the *real causal efficacy of the entities* – psychological functions – that it studies and the mechanisms of which they are part.

The methodology of psychological investigation through functional analysis is archetypally scientific. The imaginary experiment with which this discussion began adheres generally to the pattern by which an experimenter manipulates an independent variable in order to record and measure the effect on an output or dependent variable. Our two examples, Marr's *Vision* and the development of the theory of spreading activation in semantic memory are equally typical of this method. More detailed models emerge from the comparison of results from experiments with different independent variables or under which the degree or direction in which that variable is manipulated have been systematically varied.

[46] Note that this concern appears to be mostly directed at "data-fitting" or "as if" hypotheses: regularities presented as explanations.

[47] A charge to which philosophical folk psychology is equally susceptible: see Introduction, section 0.3.

All that said, the concern about the scientific status of psychology persists in some quarters. Since much of the forthcoming examination of the way psychology treats the target phenomena rests on the assumption that folk psychology, if true, should inform those investigations, I would like to take a moment to consider *three alternative ways to view psychology and what impact each would have on this contention*.

a) Psychology is **not** scientific but is a *hermeneutic* endeavour (Rennie 2012; Terwee 2012). Rather than uncovering causes, psychologists seek to catalogue, describe and rigorously define our common-sense concepts of the mind together with the regularities that pertain between these on the one hand and behaviour or further mental states on the other.
b) Psychology is **not** scientific, in virtue, for example, of the unfalsifiability of its theoretic commitments; however, psychologists either *believe* that their endeavour is scientific or seek to present it to the world as such. They avail themselves of a version of the scientific method in order to bolster its claim to scientific status – and in so doing *present* functional analyses and decompositions as *causal* explanations.
c) Psychology **is** scientific and seeks to identify causal functional relationships by means of the functional analysis of cognitive or behavioural regularities into their underlying sub-personal mechanisms.

These three possibilities are not exhaustive. They are viable alternatives and probably mutually exclusive – each seems to contradict the others. Although I would admit they are paraphrases, all have been levelled at psychology at one time or another. In the course of this chapter I have supported option c), which seems to me to best capture the way that the examples offered so far work, along with those that will appear in the next two chapters. However, I acknowledge that some might prefer to characterise psychology according to a) or b) (review: Dienes 2008) and some psychologists would claim that a) approximates their approach (Messer et al. 1988; Richardson and Fowers 2010; Sugarman and Martin 2010). However, it would not matter to this project if they are correct.

If a) is true then its explanatory aims are somewhat at odds with those of philosophical folk psychology, which *does* seek an account of the causal antecedents of action choice and regards interpersonal understanding as a species of causal understanding. We would also expect, if philosophical folk psychology is correct about the ubiquity of propositional

attitudes, that a hermeneutic philosophy would count the attitudes and their psychological formation as a significant set of "common sense concepts of the mind" and so to feature prominently in its "catalogue" of descriptions.

If b) is an accurate description of the psychological project (albeit sceptical to the point of being dismissive) then we would expect that, in pursuit of the *appearance* of scientific rigour, psychologists would help themselves to propositional attitude terms such as "beliefs" and "desires" either because these are the *de-facto* designators of the causes of action, or because they pick out features of a psychological regularity to be explained by means of analysis.

To make this explicit, *if* b) or c) are the best descriptions of psychology *and* philosophical folk psychology is true, then there are two possibilities: *either*

i) the "belief-desire law" is a psychological regularity, an *explanandum* demanding investigation, in which case we would expect to see this law or its derivatives featuring as the starting point of scientific (or even pseudo-scientific) psychological investigations *or*

ii) the terms "belief" and "desire" pick out functional entities in the mechanisms through which behaviour, action choice or interpersonal understanding occur; in which case they ought to feature in the *explanantia* of those capacities.

Given this, the examination of how contemporary scientific psychology treats the phenomena of action-choice and interpersonal understanding (in terms of action-explanation) in the next two chapters is *unaffected* by the stance that one chooses to take regarding the scientific status of psychology.

1.7 Chapter Summary

It is insufficient for explanations in scientific psychology to stop at regularities (section 1.1). Even if the regularities are between certain psychological "states" and action choices or subsequent behaviour, as is apparently suggested by the belief-desire law, then such regularities remain in need of further elucidation if they are to be regarded as genuine explanations (rather than mere "as if" statistical correlations). Investigations in scientific psychology thus do not proceed by the formulation of "laws" or "law-like regularities" but regard regularities as explananda to be investigated further.

The process by which scientific psychology investigates regularities is one of **functional analysis (**1.1-1.2). This seeks to uncover systematic relationships between ever more simple functions that interact to produce more complex, personal-level psychological functions. The objective of functional analysis is to uncover the mechanisms through which complex psychological operations are built up from these interactions. Illustrations of the process of functional analysis at work (1.3-1.5) show how central this has become to the development of psychological explanations.

The "belief-desire law", if true, would be just the sort of regularity that scientific psychology ought to take as the object of analysis. If its components, "belief" and "desire" pick out basic functions in the generation of human action, we would expect these terms or their referents to feature at some stage in the process of functional analysis. The next two chapters will examine scientific approaches to two of the human experiences that the belief-desire law is suggested to "explain – action choice and interpersonal understanding – and show that these are developed without reference to this formulation or to its components. It makes no difference what view one takes of the scientific status of psychology (1.6). If the belief-desire law works in the way that is claimed for it, then it *ought* to feature in the dominant psychological approaches to these fields of study.

2 Chapter Two:
Action Choice, Judgement and Decision-Making

Abstract: This chapter examines both the normative ideals of judgement, decision-making and action choice and the evidence from scientific psychology of the ways that such choices are actually made. The normative ***expected utility decision theory*** *is closely allied to the belief-desire model of philosophical folk psychology and so suffers from similar limitations. Not least among these is the empirical observation that people are prone to a number of cognitive illusions and biases, especially when it comes to the calculation of probabilities – which is essential to expected utility. Evidence from the circumstances under which people make choices, combined with experimental investigations into decision strategies and the computational functions that underlie them (some of which are described) suggest that people use* ***heuristics*** *to arrive at "good enough" inferences. Further evidence suggests that many – perhaps most – of our decisions are made subpersonally, out of conscious awareness and are* ***automatic****. Some contemporary cognitive psychologists posit two types of process that facilitate decision-making in the real world.*

> As far as ordinary life is concerned, the chance for action would frequently pass us by if we waited until we could free ourselves from our doubts, and so we are often compelled to accept what is merely probable. From time to time we may even have to make a choice between two alternatives, even though it is not apparent that one of the two is more probable than the other.
>
> Descartes, *Principles of Philosophy,* (1644) Article 3

2.1 Expected Utility Decision Theory: *"Homo Economicus"*

According to the belief-desire law, people will *tend (all things being equal) to do whatever they believe will bring about the fulfilment of their desires*.

If this is true, then the basis of the scientific investigation of judgement and decision making would be to explain, firstly, how this tendency arises from our psychological make-up – including how specific beliefs and desires can act upon our preferences with sufficient force

to bring the regularity about and, conversely, what factors are at work on those occasions that beliefs and desires are insufficient to bring about the action-choice predicted by the belief-desire law. Secondly, they would need to focus on the process of **belief formation** and how beliefs are generated in relation to a motivating desire. The regularity described by the belief-desire law would be central to these investigations – featuring either as the *explanandum* of a psychological theory or at some stage in the development of an *explanans* through functional analysis.

As we shall see in this chapter, however, the contemporary psychology of judgement and decision making is a vibrant area of investigation. It entails discussion not only of the origins of action and action-choice, but crosses into the philosopher's traditional territory in considering the nature – and even the existence – of human rationality (Hardman 2009: 4-5). Despite addressing broadly similar phenomena to philosophical folk psychology, the scientific approaches to judgement and decision making that I consider in this chapter make *no use* of the belief-desire law and little mention of "belief" or "desire".

We should begin, however, with "the most widely accepted principle of normative decision making" (Speekenbrink and Shanks 2013: 682). This does have much in common with that picture and which has been influential in, for example, economics, sociology and political science (Oaksford et al. 2012: 140).

Since at least the 18th century, the dominant theory of judgement and decision making has presumed that rational agents seek to maximise their *utility*[48]. The enlightenment philosopher and economic theorist Adam Smith was among the first to put the idea in formal terms when he claimed that, as "economic actors[49]", people act in pursuit of the greatest satisfaction of their material enrichment at the expenditure of minimum cost in terms of material resources, time and effort (A. Smith 1776/1986). Contemporary expected utility is a refinement of Smith's observation. It suggests that individual agents choose courses of action on the basis of a probabilistic judgement of the likelihood that a particular action will result in a given outcome together with a further judgement of the relative *utility*, or contribution to the individual's wellbeing, associated with that outcome. Dominant in Economics, this model, the **Expected Utility Decision Theory (EUDT)**, has given rise to

[48] Or at least to pursue goals that they *believe* will maximize their utility.

[49] Whether as consumers, producers, both or as regulators.

an enormous volume of scholarship – overview e.g., Speekenbrink and Shanks (2013). The essence of the theory can be stated in a single sentence:

> In choosing how to act, people will select that course that they determine has the greatest probability of maximising **utility**.

This simple statement obscures a great deal of complexity. Before I deal with some of this, however it should be noted that contrary to a common criticism of the model (Speekenbrink and Shanks 2013), EUDT does not imply **selfishness** or preclude **altruism**. "Utility" is a term of value. The degree to which an individual's actions are entirely self-serving (at one end of the scale) or entirely altruistic (at the other) is determined by *how much utility the agent attaches to altruism or to self-interest*. All that matters is that once the agent has settled on a way to calculate a value for utility they will act in the way anticipated to be most likely to maximise that value.

The first modern treatment of EUDT appears in Von Neumann and Morgenstern (1947, 2007). Their conception was of a normative account of decision-making. EUDT was offered as a template for how an *ideal decider, under ideal conditions and in possession of optimum levels of information, ought to choose*. In support of the theory they suggested that EUDT should be understood in terms of the application of a set of **axioms** of ideal decision making. Plous (1993: 81-82) sets out six axioms[50] as follows:

> **Ordering:** It must be possible to compare and rank options
>
> **Dominance:** The chosen option must never be outranked or "dominated" by an alternative
>
> **Cancellation:** Identical outcomes are disregarded in making a choice – only outcomes that differ are relevant to the decision.
>
> **Transitivity:** If outcome A is judged better than outcome B and B better than C then A is to be preferred over C.

[50] Speekenbrink and Shanks (2013: 684), list the four most important as completeness, transitivity, independence, and continuity – incorporating similar requirements to Plous' list into fewer axioms.

Continuity: If the probability of the best possible outcome far outweighs the probability of the worst possible outcome, the decider should always prefer this gamble to a *certain* "middle-ranked" outcome[51].

Invariance: the mode, order or style in which options are presented should have no bearing on the choice made since the mathematical calculation of expected utility is unaffected by such concerns.

According to Von Neumann and Morgenstern (1947, 2007) violation of any axiom would lead to a *failure to maximise utility* and so would not meet the central normative stricture of the theory – *do whatever it is that you believe has the best chance of maximising your utility*.

Most of the time, agents are confronted with a choice of several potential actions, each of which has the potential to lead to any one of a number of outcomes. Formalised versions of EUDT (Plous 1993 Ch. 7) maintain that decision-making with respect to action choice follows a series of steps:

i) Choose an optional course of action to examine.
ii) Posit the possible outcomes for that course of action.
iii) For each of these potential outcomes, calculate the *probability* that it will result from the examined course action.
iv) For each of these potential outcomes, calculate the *utility* it would deliver if realised.
v) Multiply the probability of each possible outcome by its utility value.
vi) Add these products together to give the *sum of probable utilities* for that action.
vii) Repeat this procedure for each considered course of action.

Some possible outcomes for any considered course of action are likely to have negative probable utilities (i.e., *risks* – deleterious potential outcomes that have a notable probability of occurring). This will proportionately reduce the sum of probable utilities; in this way the risk of any given course of action is given its appropriate weight.

The normatively best course of action will be the one with the greatest sum of probable utilities. This formulation measures both how advantageous or beneficial the potential

[51] According to Plous, this injunction is motivated by the need to **maximize utility** whenever possible. Thus a risky higher return has greater normative force than an assured one in the mid-range.

outcomes of an action choice are, but also the *probability* that the action will lead to a beneficial (high utility) outcome. The formula for calculation the sum of expected utilities of an uncertain choice can be written:

$$\sum \Pr(i)\,\mathrm{U(i)}$$

(Oaksford et al. 2012: 139)

Where *i* is any possible outcome of a given choice. *Pr(i)* is the calculated probability of that outcome and *U(i)* is the calculated utility for that outcome. *P(i)xU(i)*, thus calculates a probabilistic likelihood that any possible outcome will yield a given utility. The optimum choice and the normatively *correct* action would then be the one with the greatest value of for the sum of the values of probability and utility for the considered potential outcomes of that action – that is, for all examined values of *i*.

To illustrate by means of an example how this calculation is supposed to work, imagine a motorist buying a new car. They have arrived (for the sake of simplicity) at a couple of options. What will determine the final selection of a particular model, under EUDT, is the assignment of values to the features of the vehicle they are considering and the probability that each choice will deliver that utility. Suppose that one of the two cars on their shortlist offers greater fuel economy and the other greater comfort. Our motorist might assign a greater value – based on personal preference – to fuel economy than to comfort. Fuel economy thus scores greater utility than comfort. At the same time, they might be sceptical of manufacturer's figures on fuel economy which would lower confidence (probability) of achieving that outcome. After taking a test drive, the driver is *more certain* of the thirstier car's comfort than he is of the ostensibly more fuel-efficient vehicle's ability to deliver that benefit. Thus, the probability that the comfortable car will deliver its best feature is deemed higher than the likelihood that the economical model will match up to its promises. We might present this:

Model A: Comfort: U=3, P=0.8 Fuel Economy: U=7, P=0.4

Model B: Comfort : U=6 P=0.8 Fuel Economy: U=5, P=0.4

For the purposes of this demonstration it is assumed that the probability that each vehicle will deliver on comfort is the same (both have been test driven) and that the probability of achieving the promised fuel economy is also the same, given equal scepticism to

manufacturer's figures. If we carry out the probabilistic utility calculations for each option, the sum for Model A comes out at (3x0.8) + (7x0.4) = 2.4 + 2.8 = **5.2**. The sum of potential utilities (Σ) for Model B would be (6x0.8) + (5x0.4) = 4.8 + 2 = **6.8**. Thus, despite having placed higher utility on fuel economy than on comfort, and even though model B offers worse fuel economy, which had been the declared priority, the *rational* motorist ought to choose Model B.

This would be the **normatively rational** choice according to EUDT. Whether this bears any relation to how decisions are made and acted upon in the real world is questionable. For example, Hardman (2009: 66) argues:

> Most economic and psychological accounts of risky decision making[52]… assume that when people make risky decisions they are trying to maximise something such as expected value or expected utility. However, economic theories fail to capture much of human decision-making behaviour.

If, as intended by Von Neumann and Morgenstern, EUDT was regarded only as a normative prescription for ideal decision making**,** then departures from the norm would not be at issue. Nevertheless, the EUDT model underpins many **descriptive** accounts of decision-making, as is made clear by Stanovich (2011: 6), who argues that *people are presumed to behave as if they were seeking to maximise utility*. This assumption is used to construct models of the political economy by sociologists and economists (Elster 1986; Hindmoor 2006). Over the medium- to long-term and averaged over a typical customer-base/national/transnational population sample, predictions work on an assumption that individual choices will be made *as if* on the basis of expected utility (Mankiw and Taylor 2014: 3-8).

Evidence, however, suggests that decision-makers frequently *violate the axioms* of EUDT – particularly the axioms of **continuity** and **invariance**. Because they do so in predictable and systematic ways, researchers in contemporary scientific psychology have been able to *analyse the underlying processes and develop new descriptive theories of judgement and decision-making*. These are discussed in the present chapter.

This is significant because The assumption that EUDT successfully explains decision making is parallel to the metaphysical commitments of the belief-desire law.

[52] The qualification "risky" entails only that the outcome of the decision matters, that something is at stake for the decider.

2.2 EUDT and Folk Psychology: The Parallels

David Lewis was in no doubt that EUDT and belief-desire psychology are deeply intertwined.

> Decision theory (at least if we omit the frills) is not esoteric science, however unfamiliar it may seem to an outsider. Rather it is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference and choice. It is the very core of our commonsense theory of persons, dissected out and elegantly systemized.
>
> Lewis (1974: 337-38)

In this section I will examine their relationship in order to set the groundwork that both are equally susceptible to empirical challenges.

To begin with, *utility* is a quantitative measure of the power of the potential outcome of an action choice to satisfy the needs and wants (desires) of the agent. Expected utility is a process assigning a probability to the chance that the outcome of a choice will satisfy a desire. It proposes that action choice is mediated by the strength of the agent's *belief* – another way of describing an assigned probability or expectation – that a given action will meet their desire.

Expected utility decision theory and belief-desire psychology are both committed to a normative model of judgement, defining how agents *ought* to arrive at their decisions to act under ideal conditions. According to belief-desire psychology, agents *ought* to choose the action that they *believe* is most likely to fulfil their *desire*. Under EUDT, individuals *ought to weigh probabilities* and choose that option that they *believe* to be most likely to maximise their utility – synonymous with **desire fulfilment**. Expected Utility Decision Theory incorporates a specification for **normatively rational belief formation** based on the assignment of probabilities.

Given this, the underlying normative rule of EUDT can be *presented in a form similar to the belief-desire law*:

> Faced with a choice of actions, an individual should [will tend to[53]] choose the one that they estimate has the best chance of maximising their own utility.

This holds wherever "utility" can be considered synonymous with "fulfilment of their most compelling desire". There is no need of the *ceteris paribus* clause of the belief-desire law because the idea that utility should be *maximised necessarily implies that no other beliefs or desires conflict with that goal*.

The principal difference between these two approaches to action choice is that EUDT *seeks to give rational choice (and, by extension, FP) a mathematical foundation*. It employs probability theory and a calculus for weighing expected outcomes.

Empirical evidence from cognitive psychology raises doubts about the applicability of either model to the choices that people make in the actual (non-ideal) world and under the real (equally imperfect) conditions that such choices are made. Given these parallels, it is reasonable to assume that *where EUDT fails to explain the real-world decision-making process, philosophical folk psychology will suffer similar weaknesses*.

The first of these to consider emerges from application of the model of rational probability calculation: Bayes' theorem.

2.3 Dynamic Probability and the Rational Bayesian

Thomas Bayes was an 18th century English cleric, mathematician and gambler. His fascination with card games led him to develop a mathematical technique to calculate probabilities for the turn of the next card from a deck. What effect does the turn of one card have on the probability that the next card will be a particular value or suit – given that one knows the number of cards and the history of the cards produced so far? Today, Bayes' theorem is central to the calculation of this kind of **dynamic probability –** how odds change in the light of new information – in fields far removed from games of chance. For example, it is used to account for how scientists' confidence in the truth of their theories is affected by each new piece of experimental or observational data (Losee 2001: 221). It is also used by proponents of Expected Utility Decision Theory to determine how initial estimations of the probability that a given course will maximize utility ought to be revised in the light of

[53] The belief-desire law is presented as if it is descriptive as well as normative, suitably *hedged* by the claim that it is a *tendency*.

new information – e.g. about the environment, the choice itself, the needs of the agent or even the interim results of the action that has been chosen (Hardman 2009: 26; Speekenbrink and Shanks 2013).

The starting point of any Bayesian calculation is the **prior probability**, often known as the **base rate**. The theorem allows us to calculate that the probability of any future occurrence in the light of a new occurrence is equal to the probability of the new occurrence (B) given the expected occurrence (A), multiplied by the prior probability (base rate) of the expected occurrence divided by the prior probability of the new occurrence. This can be represented as:

$$\mathbf{P(A/B) = \frac{P(B/A) \text{ x } P(A)}{P(B)}}$$

Behind this very brief introduction to Bayes' theorem lies a great deal of complexity and some of the essentials of the way that mathematicians deal with probability. It will suffice for my present purposes, however, if we take note of two features of Bayesian probability:

a) The accurate (or at least closely approximate) calculation of base rates is essential to the application of this theory to the calculation of dynamic probabilities.
b) EUTD and Folk Psychology both assume that people use a version of the Bayesian calculus to estimate the likelihood that their actions will bring about a given result.

2.4 Bayesian Missteps, Framing and Biases

These assumptions would mean that if people turn out to be poor at applying Bayes' theorem or at estimating base rates, the prospects for the usefulness of EUTD (and, analogously, the belief-desire law) are not good. Investigations in cognitive psychology strongly suggest that we are all pretty poor Bayesians.

For example, take the well-documented effect known as the **base-rate fallacy** (Birnbaum 2004; Tversky and Kahneman 1974). In their famous exposition, Tversky and Kahneman offered their respondents the following scenario:

> Jack is a 45-year old man. He is married and has four children. He is generally conservative, careful and ambitious. He shows no interest in political and social

> issues and spends most of his free time on his many hobbies, which include home carpentry, sailing and numerical puzzles.
>
> (Tversky and Kahneman 1974)

The subjects were asked to assess the probabilities that "Jack", as described in the scenario, was an engineer (on the one hand) or a lawyer (on the other). To "guide" their estimates, each of the respondents was told that the description of Jack was taken at random from 100 such personality sketches. Depending on which experimental condition they were in, participants were given the following additional information:

i) Half of the respondents were told that the hundred sketches were taken from a sample which included 70 lawyers and 30 engineers.
ii) The other half were told that the proportions were reversed – 70 engineers and 30 lawyers.

According to the *base rate*, that is the starting probability before additional information – in this case the content of the personality sketch – was considered, *the chances that a sketch drawn at random from the sample would be that of an engineer would be 0.3 in condition (i) and 0.7 in condition (ii)*. If the base rate was taken into account at all we would expect to see a significant difference between the two groups. No such variance was recorded. *Respondents under condition (i) and condition (ii) estimated the probability that Jack was an engineer at 0.9*. The base rate was apparently disregarded in favour of a judgement based on the only other information about "Jack" that the subjects had to work with: the personality sketch.

Birnbaum (2004) notes that similar effects in which base rates are neglected in favour of other, more compelling information have been shown consistently since Tversky and Kahneman (1974) first described it as the **base rate fallacy.** However, Birnbaum challenges its characterisation as a fallacy: **base rate neglect** might be more accurate, since, he claims, all of the investigations that have shown similar results have involved "very restricted" experimental conditions, rendering the conclusion "quite fragile" (Birnbaum 2004: 55). The lack of evidence of a robust effect of base rates is not sufficient evidence to show that base rates have *no* effect, he argues.

Nonetheless, Birnbaum is quite convinced that people fall some way short of being ideal Bayesian probability estimators, as illustrated by an example that he offers:

> Suppose there is a disease that infects one person in 1000, completely at random. Suppose there is a blood test for this disease that yields a "positive" test result in 99.5% of cases of the disease and is a false "positive" in only 0.5% of those without the disease. If a person tests "positive", what is the probability that he or she has the disease? The solution, according to Bayes' theorem, *may seem surprising*.
>
> (Birnbaum 2004: 44 – emphasis added)

Surprising, in Birnbaum's word, because despite the 99.5% accuracy of the test, the Bayesian probability that the subject of the positive test actually has the disease is *just 0.166'*. Although this is still more than 166 times greater than the probability that the subject is sick prior to the incorporation of the test information (1/1000 or 0.001) Birnbaum's contention (which seems plausible) is that this will strike most people as very low given that they have registered "positive" on a test that has only a 0.05 probability (0.5 %) of being wrong.

That the result of applying Bayesian probability to these numbers is *surprising* also suggests that we disregard base rates when making probability judgements. Although sceptical of the contention that the evidence shows us to have *no* conception of base rates, Birnbaum proposes a more modest claim. For him, the results suggest that although sensitive to probabilities, we make decisions in a way that is not constrained by Bayes' theorem – despite the fact that this model is central to the Expected Utility Decision Theory. Often, other influences override base rates and the Bayesian calculus and lead us "wilfully" to neglect both.

Base rate neglect is not the only way in which people tend to diverge from the Bayesian ideal when judging probability on the basis of new information. Take, for example, the **Monty Hall Dilemma**[54] (De Neys and Verschueren 2006). The basic description of this puzzle, upon which there have been countless variations (Burns and Wieth 2004; Camerer 1995), begins by inviting the participant to imagine that they are taking part in the final phase of a television game show. The "host" shows them three closed doors, labelled 1, 2 and 3. They are told that behind one of the doors is a considerable sum of money – let's say

[54] Monty Hall was a US television game-show host of the 1970s. It has not been established that this demonstration figured in any show that he presented; it is perhaps just the "sort of thing" that he *might* have featured.

$64,000. Behind each of the other doors is either nothing at all, or, in some versions, a booby prize – such as a live goat!

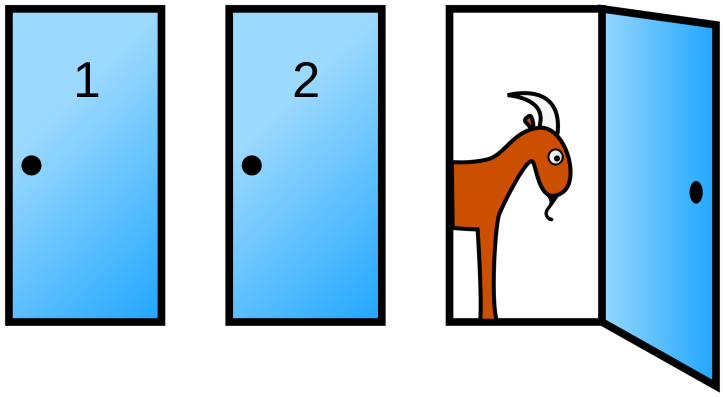

Fig. 2.1 The Monty Hall Problem

The subject or "contestant" is invited to choose one of the doors and told that they will win whatever prize is revealed when the door of their choice is opened – a hungry ruminant or a pile of cash. The contestant/subject duly makes their choice and announces which it is to be – 1, 2, or 3. When this scenario has been run as a psychological experiment there is, as would be expected, a random distribution of choices between the three (Burns and Wieth 2004).

In keeping with the ritual cruelty of the game show genre, however, the host/experimenter then introduces another factor. One of the two doors that the participant did not choose is opened and reveals – a goat!

The contestant/subject is then given a choice: stick with their original choice or switch to the other door – which they now know conceals *either* a second goat *or* $64,000.

If you are the contestant, do you stick or switch?

In experimental settings most – in some instances as many as 85% (Burns and Wieth 2004) – stick with their original choice. Many report that this *feels intuitively* to be the right choice. By the rules of probability, however, *it is the wrong choice. To maximise their chances of walking away with the money, rather than the goat, the contestant should always switch*.

To understand why, imagine that the host opens door 3, revealing a goat. When offered the "stick or switch" choice, there are three possibilities:

i) The contestant's first choice, *door 1*, hides the other goat. The host has opened the only remaining non-winning door so switching to door 2 is *certain* to result in success.

ii) The contestant's first choice was *door 2* but this was in fact the other booby-prize door. We now know that there is a goat behind door 3 so switching to door 1 is *guaranteed* to be successful.

iii) The contestant chose the winning door (whether 1 or 2) the first time and so will *win* if they *stick* to their original choice.

In two out of three of these scenarios, switching is successful. In only one out of the three does the subject win the money. Switching has a probability of success of 0.666', whereas their chances if they stick are just 0.333'. Yet when asked what effect the host's opening of the other door had had on the probability that their original choice was the right one, most respondents, according to Burns and Wieth (2004), claimed that it had increased from 0.333' (one in three) to 0.5 (one in two). In fact, *the probability of their original choice being the right one is still one in three*: the likelihood that the *other* remaining closed door is the best choice is now *twice as good*, given that we now know that door 3 is a losing choice.

Even when this reasoning is explained, many subjects insist that switching *feels* counterintuitive (Eysenck and Keane 2010: 461). This "feeling" is sufficiently strong to override good probability reasoning. Of course, a contestant who sticks will still win the money one time in three. All that is certain is that when equipped with a knowledge of the way that probabilistic reasoning works, *switching is the rational choice*. That it is the minority choice indicates that most of us are poor estimators of probability.

For an illustration of another way that judgements diverge from the rational norm, imagine a patient facing a diagnosis of *lung cancer*. An oncologist suggests that there are two treatment options: radiation therapy and surgery. He presents the options and the statistics on their outcomes as follows:

> Surgery: Of 100 people having surgery, 90 live through the post-operative period, 68 are alive at the end of the first year and 34 are alive at the end of five years.

> Radiation Therapy: Of 100 people having radiation therapy, *all* live through the treatment, 77 are alive at the end of one year and 22 are alive at the end of five years.

For both options, the figures are presented, or **framed**, in terms of survival.

Would it make any difference to the patient's decision if the same statistics were instead framed in terms of mortality?

> Surgery: Of 100 people having surgery 10 die during surgery or the post-operative period, 32 die by the end of the first year and 66 die by the end of five years.
>
> Radiation Therapy: Of 100 people having radiation therapy, none die during treatment, 23 die by the end of one year and 78 die by the end of five years.

It is important to note that the numbers under each frame, the proportion of a sample of patients who live or die within a particular time period, are exactly the same. The only difference is whether attention is drawn to the *number of deaths or to the number of people who survive*. McNeil et al. (1982) carried out a study based on this scenario in which groups of medical students, business students and hospital patients were asked to imagine themselves faced with such a choice. Half of the subjects within each group were presented with the information under the survival frame and half under the mortality frame. Overall, across the three groups, 18% of respondents favoured radiation therapy when they were given the survival frame, whereas 44% of those exposed to the same figures under the mortality frame said that they would choose the radiation therapy option. These figures seemed stable across all three groups of respondents – medical students, patients and business school students all found radiation therapy more than twice as compelling if they were informed of the figures via the mortality frame.

Tversky and Kahneman (1985, 1986) pointed out that these results represent a violation of the **invariance principle** of EUDT, stated on p74, above as:

> The mode, order or style in which options are presented should have no bearing on the choice made since the mathematical calculation of expected utility is unaffected by such concerns.
>
> (Plous 1993: 82)

Contrary to that normative constraint, these findings suggest that the framing of the figures has a marked effect on the choice that people would make. Whatever strategy these subjects were using to make their choice it was not on the basis of expected utility since the outcomes and the probabilities of each outcome are the same regardless of the way the figures are presented. The true EUDT decider would make the same choice, regardless of frame. Framing effects have been recorded according to various ways that information is presented, including the order in which it is presented (Gilovich et al. 2006: 397)[55].

Framing effects have direct bearing on the idea that our choices are determined by whatever we *believe* will bring about the fulfilment of our *desires*. Assuming that the respondents to McNeil et al in both conditions (survival or mortality frame) shared the same *desire* to survive for as long as possible following their imaginary cancer treatment, then what they *believe* about the way to achieve that goal is the same under each presentation of the figures. Again, *some other consideration takes precedence over the belief-desire law in choosing a course of action*.

Cognitive psychology has also uncovered a number of other *systematic* (reproducible and predictable) **biases** in the ways that people decide which action to take. One classic demonstration of a bias is the **Wason Selection Task** as reported in Wason and Johnson-Laird (1970).

Subjects are presented with four cards lying on a table. On the visible side of each card is printed a single letter or number:

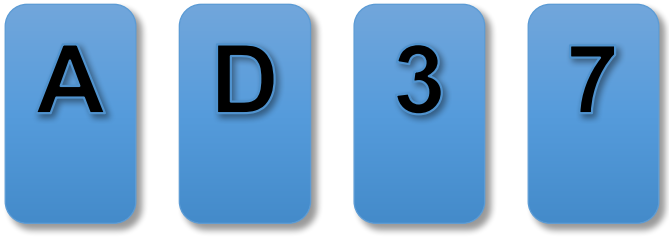

Fig. 2.2: The Wason Card Test

[55] For further examples of framing effects in different domains, see E.J. Johnson et. al. (1993), E.J. Johnson & Goldstein (2003) and the essays in the collection edited by Lichtenstein & Slovic (2006).

The participant is told that it is certainly true that every card has a number on one side and a letter on the other. They are also given a further rule about the cards which, they are told, might be true or false:

> Any card which has the letter A on one side has the number 3 on its reverse.

The participant's task is to determine whether or not this statement is true by turning over the minimum number of cards. Which cards should they turn over first?

In the original demonstration (Wason and Johnson-Laird 1970) and in numerous subsequent replications (Evans 2004; Eysenck and Keane 2010: 542-43) the majority of subjects choose to turn over either the *A card* alone or the *A* and the *3* cards. This is not, however, sufficient to determine the truth of the statement. If the *A* card is turned over and does have a *3* on its reverse, this proves only that the statement is true for one instance. We cannot extrapolate from this that it is universally true. Of course, turning over the *A* card alone and finding a number *5* on the reverse would be sufficient to *falsify* the statement. Suppose that we do find a number *3* on the other side of the *A card*: subsequently turning over the remaining original *3* cannot decide the truth or falsity of the statement either. If an *A* is printed on the back, we are no better off than when we found a *3* on the reverse of the *A card* – we now have two instances of the rule but still nothing that allows us to deduce whether it is generally true. If the reverse of the *3 card* is the letter *B*, this does not disprove the statement either: the contention was that any A has a 3 on its reverse – no general claim was made about all cards marked *3*.

The best course of action is to turn over the *A* first and then, if this does not falsify the statement (by having any number other than a *3* on its other side), to turn over the *7 card*, which offers another opportunity to falsify the proposition by having an *A* on its reverse.

In their original interpretation of this result, Wason and Johnson-Laird (1970) theorised that *people have a tendency to assign greater weight to evidence that would support or confirm the hypothesis that they are trying to examine rather than that which would falsify it* – they exhibit **confirmation bias.** Hardman (2009: 94, 98) reviews research that suggests an alternative interpretation: subjects show a marked preference for *matching* their selections to the options that are mentioned in the original question (Evans and Lynch 1973; Evans 2004).

For the purposes of the present thesis, the precise underlying mechanism that produces this effect is irrelevant. That such effects seem robust and reproducible, suggests unequivocally that people's reasoning tends to diverge systematically from the optimal deductive strategy.

2.5 Fast and Frugal Heuristics: "*Homo Heuristicus"?*

Growing evidence, both from the features of *persistent systematic biases and cognitive illusions* (Kahneman 2011; Tversky and Kahneman 1974) and from *experimental investigations of performance on judgement and decision-making tasks* (Gigerenzer 2008; Gigerenzer and Gaissmaier 2011) suggests that much real-world rationality depends on the employment of **heuristics**. The word "heuristic" derives from the Greek for "serving to discover" (Gigerenzer and Brighton 2011: 3).

Work on the use of heuristics in decision making began with the heuristics and biases research pioneered by (Tversky and Kahneman 1974). Work focussed on the positive benefits of heuristic reasoning has been pursued principally by the Fast and Frugal Heuristics movement, led by the Centre for Adaptive Behaviour and Cognition or ABC group (Gigerenzer and Todd 1999b; Gigerenzer et al. 2011). For the debate between these two approaches see Gigerenzer (1991, 1996) and Kahneman and Tversky (1996).

Heuristic "rules of thumb" *do not entail a particular outcome*, in contrast to normative models of inference. Instead, they deliver judgements that are:

i) **advantageous** such that the mechanisms delivering them would be potentially selectable by evolution through natural selection (Gigerenzer and Todd 1999a: 32-33) and
ii) "**good enough"** (Gigerenzer 2008: 81) to allow the individual to negotiate their world successfully.

"Good enough" implies that these judgements might not be *optimal* – i.e., the best possible – but that they are *successful outcomes* by the ecological standard of allowing the organism to survive and thrive.

Arguing for scientific realism, Putnam (1982) suggested that *unless scientific theories are true, then their explanatory and predictive success would be some kind of miracle*. Faced with evidence of systematic deviations from *normatively correct* decision-making, it is tempting to make use of a similar **no miracles argument**: *how can our species have been*

so successful, have achieved so much, unless we are, mostly, rational? This assumes, however, that the *normative requirements define what "rational" means*. According to cognitive psychologists researching the heuristic basis of judgement and decision-making (Gigerenzer and Brighton 2011) there is a better way to describe rationality of the kind that has bestowed our advantages.

To be fully realised, a decision made under Expected Utility Decision Theory needs to *maximise utility*. Ideally, this would entail taking into account all of the possible outcomes of a choice, accurately assigning a utility value to each of these and accurately calculating the probability that a given action choice will result in a specific outcome. The problem with this is that humans have *limited time, computational resources and information* on which to base a decision. This implies that *every decision entails an additional decision* – when to *stop the process*:[56] that is, the decision of when further investment of time and effort in acquiring and processing additional information will not produce a sufficiently better decision to make that investment worthwhile. The stop-decision must, if the optimisation requirement is to be respected, itself be **optimal**. *This leads to a potential regress* – we need to make an optimal stop decision in order to justify the decision to stop the investment in the principal decision, which would, in turn need to be optimised – and so on *ad infinitum*. This regress is one of the limitations of so-called **optimization under constraints** models (Gigerenzer and Todd 1999a: 10).

Simon (1955) described the kinds of processes through which we make decisions under impoverished conditions of information, time and capacity as **bounded rationality**. He theorised that individuals operating under these constraints consider their options more or less sequentially and then *stop looking when they find one that minimally meets their immediate need*[57]. Simon described this kind of stop-decision based on need-fulfilment (rather than optimisation) **satisficing**. According to Gigerenzer and Todd (1999a: 12) there has been a tendency to regard bounded rationality as synonymous with optimisation under constraints – for example, the economist Sargent (1993). This is incorrect. As Simon (1991: 35) writes:

> Rationality need not be optimization, and bounds need not be constraints.

[56] Cf. the frame problem in artificial intelligence (McCarthy and Hayes, 1981).

[57] Note that this contrasts with the "optimal" or "maximal" target of normative models.

To illustrate the contrast between normative and ecological rationality, consider this picture of a decision-making process, adapted from Gigerenzer (2008: 21-22). An outfielder in a ball game – such as baseball or cricket – attempts to catch a high ball. The traditional view of the computational processes involved might be rendered:

> He behaves as if he had solved a set of differential equations in predicting trajectory of the ball. He may neither and know nor care what a differential equation is, but this does not affect his skill with the ball. At some subconscious level, something functionally equivalent to the mathematical calculations is going on. Similarly, when a man makes a difficult decision, after weighing up all the pros and cons, and all the consequences of the decision that he can imagine, he is doing the functional equivalent of a large 'weighted sum' calculation, such as a computer might perform.
>
> (Dawkins 1989: 96)

An *optimised under constraints* account involves the same series of computations, with an added stopping-rule: i.e., stop search as soon as the costs of further search outweigh the benefits (Gigerenzer and Todd 1999a: 10). Once all this computation has been performed, assuming that, by some unspecified method, the infinite regress can be circumvented, the fielder will know where the ball will come to ground and can run to that location, make the catch and enjoy the accolade of the crowd. Alternatively:

> Fix your gaze on the ball, start running, and adjust your running speed so that the angle of gaze remains constant.
>
> (Gigerenzer and Brighton 2011: 3)

This is the **gaze heuristic**; it is one of a *number of simple rules of thumb that experiments have found are used by skilled fielders in the process of catching a ball* (McLeod and Dienes 1996). Like many heuristics, these are employed unconsciously and automatically, although there is no principled reason why a heuristic, once stated, cannot be employed consciously. A definition of a heuristic might be:

> A strategy, conscious or unconscious, that searches for minimal information and consists of building blocks that exploit evolved capacities and environmental structures.

(Gigerenzer 2008: 22)

An alternative description of the heuristic decision-making process is offered by Plous (1993: 107):

> When people are faced with a complicated judgement or decision, they often simplify the task by relying on heuristics, or general rules of thumb. In many cases, the shortcuts yield very close approximations to the optimal answers suggested by normative theories. In certain situations, though, heuristics lead to predictable biases and inconsistencies.

Although these "biases and inconsistencies" initially drew researchers' attention to the use of heuristics (Tversky and Kahneman 1973, 1974), other researchers have focussed on the ecological value of **fast and frugal** (Gigerenzer and Gaissmaier 2011) judgements derived from the use of heuristics play in decision making. **Fast**, because they *allow decisions to be made under time constraints* and **frugal** because they use *just enough information and computational resources to deliver a good-enough choice* (Gigerenzer and Todd 1999a; Gigerenzer and Goldstein 2011).

When Gigerenzer writes in the quotation above that heuristics exploit "evolved capacities and environmental strategies" he refers to the presumption that heuristics are not generally thought to require dedicated processing resources – there is no "heuristic module". Models of heuristics assume that they exploit other, more basic features of cognition, such as **recognition** (Gigerenzer and Gaissmaier 2011). Features of cognition that have evolved under a variety of demands are available for heuristic decision making.

Individual heuristics, once acquired, are selected for use on the basis of *cues* drawn from the task to be performed. Four possible strategies for the matching of a given heuristic to a specific task have been suggested (Gigerenzer and Gaissmaier 2011). Some heuristics may be hardwired by evolution, such as the "extension of objects in three dimensional space". Others may be individually learned by a process of trial and error. The third strategy might involve social learning, either being taught rules by parents and peers or, less formally, noticing how others approach problems and applying their heuristics whenever a remembered cue suggests it might be appropriate. Finally:

> …the content of individual memory determines in the first place which heuristics can be used, and some heuristics' very applicability appears to be

> correlated with their "ecological rationality". For instance, the *fluency heuristic* is most likely to be applicable in situations where it is also likely to succeed.
>
> (Gigerenzer and Gaissmaier 2011: 456)

This **fluency heuristic** is a typical example of a simple rule for inferring which of two alternatives is the better choice even when both are recognised.[58]The rule can be stated as:

> If both alternatives are recognized but one is recognized faster, then infer that this alternative has the higher value with respect to the criterion.
>
> (Ibid.: 462)

An example would be if you are asked which of a pair of vaguely familiar cities has the largest population. Use of the fluency heuristic suggests that whichever name triggers the more fluent recognition is likely to be the better known – and so, according to ecological rationality, to have the larger population The fluency heuristic is related to the **recognition heuristic** (Pachur and Hertwig 2011), which would be appropriate in this example if only one of the alternatives was recognised. Experiments have suggested that heuristically derived estimates of two cities' relative populations is more likely to produce a correct result than local knowledge: German students, for example, do better at estimating the relative size of American cities – presumably on the basis of which of the two is known to them – than do American students who presumably know both (Gigerenzer 2008: 25; Goldstein and Gigerenzer 1999).

Schooler and Hertwig (2011) suggest that the process of *forgetting* plays an essential role in clearing away superfluous information and "aids discrimination between the objects' recognition speeds. In contrast, models of judgement based on optimisation, under which as much information as possible is required for the best judgements, do not leave room for the process of forgetting, despite the fact that many contemporary cognitive psychologists suggest that forgetting (or **retrieval inhibition**) is an essential part of the management of memory (Bjork 1989; Schilling et al. 2014).

There are four stages to the heuristics research programme:

[58] And thus the *recognition heuristic* – "prefer the one you know" – is unavailable.

> a) designing computational models of candidate simple heuristics, b) analysing the environmental structures in which they perform well, c) testing their performance in real-world environments and d) determining whether and when people really use these heuristics. The results of the investigatory stages (b), (c) and (d) can be used to inform the initial theorizing of stage (a).
>
> (Gigerenzer and Todd 1999a: 16)

Individual heuristics are differentiated by three factors established through that process:

> A model of a heuristic specifies (i) a process rule, (ii) the capacities that the rule exploits to be simple and (iii) the kinds of problems the heuristic can solve, that is the structures of environment in which it is successful.
>
> (Gigerenzer 2008: 24)

Theorists in the Fast and Frugal Heuristics programme categorise heuristics according to *the set of cues over which they operate, their application and the computational characteristics that they exploit* (Gigerenzer and Todd 1999a) rather than by the systematic biases and errors (Tversky and Kahneman 1974) or deviation from "rational norms" that they produce – although these, being systematic, can be useful clues as to the structure of the underlying computational processes (Gigerenzer and Gaissmaier 2011). Three heuristics were identified in Tversky and Kahneman's 1974 paper: *representativeness,* in which the cue to the choice of category under which to place an individual or item is how closely it matches a stereotype or prototype of that category (the case of judging whether "Jack" is an engineer or lawyer, above, would be such an example), *availability*, under which a choice is ranked more likely according to how easily it is retrieved from memory, and *adjustment from an anchor,* under which the starting point of a decision or calculation is the cue from which a guess about magnitude of the final value can be constructed – which is why people given limited time (typically five seconds) will usually estimate the sum of *11+27+48+56+90* to be markedly lower than they would if asked to estimate *90+56+48+27+11* (Mussweiler et al. 2004; Tversky and Kahneman 1974).

One of the key tasks that remains for the heuristic research programme is to identify which and how many individual heuristics make up the *adaptive toolbox* (Gigerenzer and Todd 1999a). This will entail analysing the real-world judgements that people make rather than

trying to fit them to a normative model or going in search of supplementary theories when they cannot be accommodated.

Keren and Teigen (2004) point out that researchers in heuristics do not offer "…a comprehensive theory that can encompass all of the heuristics under one framework…" and doubt that such an overarching theory is even feasible, given that the heuristics which have been suggested appear to make use of very different input data (including perceptions) and cognitive (computational) capacities. However, among the leading lights of research into heuristic decision making, neither Kahneman (2011) nor Gigerenzer (2008) claim that identifying those occasions on which heuristics are used to make judgements or to facilitate decision making will provide a single, unitary description of the mechanisms involved. There is unlikely to be a "heuristic system" any more than there is a discrete "non-heuristic" reasoning pathway. Both agree that heuristics play a significant role in our judgement processes – much more marked than had previously been suspected – and that understanding how they work and how they generate inferences is essential to understanding human reasoning.

Heuristics can be applied with or without conscious awareness. For example, recognition heuristics can lead us to make probability judgements without our being aware of how we have come to the determination that an event is more (or less) likely (Pachur and Hertwig 2011). On the other hand, air pilots are taught a heuristic for avoiding mid-air collisions – "*if the relative bearing of another aircraft in your sight remains constant, then a danger of collision exists"*. Likewise, soldiers working as battlefield medics are given the heuristic "*attend to silent casualties first*" which, although it goes against our instinctive reaction to give immediate attention to someone screaming in pain, respects the medical fact that somebody who can cry out is exhibiting cardio-pulmonary function.[59] Through learning, practice and repetition, even consciously learned heuristics can become incorporated into an individual's automatic decision-making routines: they are implemented without being consciously selected and without awareness of the procedure(s) (search rule, stop rule, decision rule) that they involve (Rieskamp and Otto 2011; Schooler and Hertwig 2011; Sykes 2006: 206-12)[60] (see next section).

[59] Both of these are drawn from my personal experience.

[60] This transition from conscious to unconscious competence (expert performance) is significant in work on the ACT-R cognitive architecture (see Anderson 1995), a detailed exposition of which is outside the the scope of the present thesis.

Gigerenzer and Brighton (2011) suggest that rather than imagining our species as *homo economicus*, the personification of the EUDT model of judgement and decision-making, it would be more helpful to see ourselves as *homo heuristicus*:

> Viewing humans as homo heuristicus challenges widely held beliefs about the nature of cognitive processing and explains why less processing can result in better inferences.
>
> (Ibid. 26)

I contend that the bounded, ecologically rational and open-ended "tool-kit" approach of the heuristics programme promises *to more closely map onto the way that real people, in the real world, make real decisions than do the assumptions of philosophical folk psychology or its sibling, Expected Utility Decision Theory*. This contention features throughout this thesis. Further investigations into the roles of heuristics will feature in Chapter 3 – where the focus is on social judgements – and in Chapter 4 – where I consider how heuristics contribute to judgements about the acceptability of narrative accounts.

Investigations into heuristics stand alongside another strand of research into judgement and decision making concerned with the range of choices that we make without being aware of the process or even, in many instances, that a choice has been made at all.

2.6 Automaticity and the New Unconscious

In an influential paper, Kihlstrom (1988) draws together research to that date that indicated "the impact of nonconscious mental structures on an individual's conscious experience, thought, and action". Contemporary research into the cognitive unconscious and automatic cognitive processes gained impetus from this paper although, as Kihlstrom points out, the observation that a great many of our judgements and decisions take place without conscious awareness or control has a long history throughout the philosophy of mind (see also Dehaene 2014: 49-52). Kihlstrom suggests that although "scientific psychology began as the study of consciousness, … Quite quickly… observations in both the laboratory and the clinic suggested that mental life is not limited to conscious experience".

However, thanks in part to the rise of **behaviourism** (Watson 1913) and more latterly because of the association of the unconscious with speculative processes suggested by Freud (Freud 1999, 2003), serious psychological research into non-conscious processes was

neglected through the middle years of the 20th century. However, as Kihlstrom (2013: 176) points out, "The new view of the unconscious … owes little to Freud".

Research into this **new unconscious** (Hassin et al. 2005) is often described under the heading **automaticity** (Uleman et al. 2007). As Moors (2013: 164) makes clear, we can identify a mental operation as automatic according either to the *mechanisms* that underlie it – arrived at by *functional analysis* – or by reference to the *features* of how it occurs. In this section I am particularly interested in *feature-based descriptions of operations concerned with judgement and decision-making*.

Automatic operations can be distinguished by four features. Here I have paraphrased those features in terms outlined by Bargh (1984) – the words in brackets are those that Bargh uses to identify the features[61]:

> *Unconsciousness (awareness):* The person making the automatic judgement is unaware of the process involved, the cues that elicit it (stimulus) or the "determining influences".
>
> *Involuntariness (intentionality):* The judgement is initiated, progressed and completed regardless of the conscious intentions of the person making the judgement.
>
> *Effortlessness (efficiency):* the decision process does not compete for cognitive resources with other processes. Even if the individual making the judgement is simultaneously carrying out another cognitive task, performance is unaffected.[62]
>
> *Autonomy (controllability):* the goals or motivations of the individual will not affect the speed or completion of the process.
>
> (after Bargh 1994)

[61] Bargh's original four, "awareness, intentionality, efficiency and control" are inconsistent in that one (efficiency) is a feature that he suggests automatic processes possess while the other three are features that they lack. He also uses "Intentionality" differently from its philosophical meaning ("aboutness"). For clarity I have used different names, all of which are positive features of automatic judgements.

[62] Confirmed by comparative response-time measurements.

Using this model to delineate automatic operations, scientists have begun to explore the circumstances under which they occur, the behaviours that result – including how these compare with consciously controlled behaviours – and the variety of stimuli that elicit automatic operations (Uleman et al. 2007). In the process, researchers have uncovered more and more aspects of our mental lives that have this character.

> The recent social-cognitive work on the automaticity of higher mental processes, such as those underlying social interaction, affect and evaluation, motivation and goal-setting, and social judgment, … has found much of an individual's complex psychological and behavioural functioning to occur without conscious choice or guidance—that is, automatically
>
> (Bargh and Ferguson 2000)

Theorists have supported this contention that complex cognitive operations occur without conscious awareness or control through experimental approaches. Not only judgement and decision-making: we might not always be aware of how we are influenced by visual stimuli, which according to Spencer (2007), is a direct challenge to the "platitudes of Folk Psychology": after all, can a visual stimulus that affects behaviour but is unavailable to conscious awareness, control or verbal report be characterised as a "belief"? Other investigations have shown that decisions can be made even when *conscious faculties are fully occupied with other tasks* (Bargh and Chartrand 1999; Bargh and Ferguson 2000; Dijksterhuis 2004).

Not only are we unaware of the processes that underlie a good many of our judgements and decisions (and are unable to stop them, redirect them or otherwise consciously direct their behavioural outcomes) but we are also unaware of the influences that trigger them and determine their outcomes (Bargh and Chartrand 1999; Shepherd 2015). Priming effects, encountered earlier in the present chapter, would be an example of this kind of influence, as when Bargh et al. (1996) famously showed that being unwittingly primed with words relating to *old age* tended to *reduce the walking speed of students*.

The evidence suggests that what we *believe* about a situation or even what outcome we would claim to *desire* might not be in the central factors steering our decision-engines that the belief-desire law appears to enshrine. Indeed, those decision engines that make up our *rationality*, in a non-normative sense, perform a variety of tasks without any need for our

conscious attention to what we believe or desire, or any ability to articulate such preferences. Being unable to articulate our decision-making processes does not preclude part of the information over which these tasks operate being beliefs or desires, *unless we insist that "belief" and "desire" pick out the general causal roles enshrined in the belief-desire law*. In order to fit automatic judgements into a belief-desire law format, we would be forced to redefine "belief" and "desire" so broadly that they become meaningless. They would need to encompass any information that directs a decision, including primes, unconscious biases and preferences, unconscious visual stimuli, implicit stereotype activations, for example as well as unattended information about bodily and environmental states (at a minimum). All of these have been shown to influence our automatic judgements and decisions (Higgins 2005; Kahneman 2011; Uleman et al. 2007; Wegner 2005).

Automaticity is thought to be an essential component in the development of skill through practice (Eysenck and Keane 2010: 193-99). A skilled athlete may make thousands of decisions in the course of a game yet all of them might exhibit the four characteristics of automatic processes (McLeod and Dienes 1996). An experienced physician might intuit that something is not right with their patient, long before scientific investigation can uncover a more accurate diagnosis (Chapman 2004). It is fair to suggest that an important difference between an expert and a novice lies in the depth and complexity of cognitive operations that can be completed without conscious control (K. E. Johnson 2013). Part of the process of achieving expertise in a particular domain can be the acquisition and habituation of a particularly apt *heuristic* to the level where it no longer requires conscious attention either to initiate or operate (Garcia-Retamero and Dhami 2011; Gigerenzer 2008: 42; Rieskamp and Hoffrage 1999).[63] One reason for the biases and divergences from normative rationality that drew Tversky and Kahneman (1974) into the investigation of heuristics in the first place is thought to be that we are unaware either that a heuristic is being used or that the outcome of a choice is heuristically derived (Uleman et al. 2007).

It has been argued that persistent divergences from Bayesian rationality such as the Monty Hall problem (above) might be due to the engagement of automatic reasoning in response to limitations of working memory capacity. Conscious attention and control requires that

[63] It has also been noted that lay people using a suitable heuristic can outperform "experts" employing more normatively rational models of decision-making – for example; Bernhard Borges et al., 'Can Ignorance Beat the Stock Market?', in Gerd Gigerenzer and Peter M. Todd (eds.), *Simple Heuristics That Make Us Smart* (Oxford: Oxford University Press, 1999), 59-72.

the factors influencing a judgement be held in working memory (De Neys and Verschueren 2006).

Beyond the feature-based delineation of automatic processes, there seems to be no clear-cut distinction between the kinds of judgements or decisions that might be made automatically and those that *must* engage conscious control (Moors 2013: 173). Each decision is likely to involve some parts of the process that are conscious and controlled and others that are automatic.

> Mental processes at the level of complexity studied by social psychologists are not exclusively automatic or exclusively controlled, but are in fact combinations of the features of each. […] a process can have some qualities of an automatic process (e.g., efficient, autonomous), while simultaneously having qualities of a controlled process as well.
>
> (Bargh 1994: 3)

Nevertheless, a great many of our day to day decisions and action choices are automatic, from the "decisions" that I am making about where to place my fingers while typing this to the number of times we chew our food before swallowing. Science would agree. Automatic judgements have been isolated and measured in how we draw causal inferences (Hassin et al. 2002) to the study of emotional responses (Feldman Barrett et al. 2007) or from how we conduct close relationships (Chen et al. 2007) to how we make decisions more generally (Bodenhausen and Todd 2010). It has even been suggested that the attribution of mental states to others – an essential part of belief-desire psychology and the central skill in so-called **Theory of Mind** – is a largely automatic process (Butterfill and Apperley 2012). See section 3.6 in the next chapter for more on the heuristics and automaticity of social inferences.

Just as researchers into automatic cognition are approaching the subject in a distinct way from Freud and his followers, admitting that much decision-making happens *sub rosa* is not a kind of **behaviourism**. Cognitive scientists and psychologists investigating automatic processes seek to uncover the processes involved by a process of functional analysis, as described in Chapter 1. It is insufficient merely to catalogue regular conjunctions of "stimulus and response" (Skinner 1974).

> We have argued further, however, that it is an error to conclude that those processes that do require the intervention and guidance of conscious or executive control processes – such as those that involve the flexible and strategic operation of working memory – are any less determined, because such processes are also caused. Therefore, the task of future cognitive and social-cognitive research should be, as Baddeley (1996)[64] and others have recently argued, the discovery and delineation of the mechanisms by which such executive processes operate.
>
> (Bargh and Ferguson 2000)

Given that so much of what is generally regarded as "cognition" can be undertaken without attention, it would seem that, *pace* Freud, attention itself is the most enigmatic aspect of our mental life (Sykes 2006). The pressing question for some researchers in cognitive psychologists (as for philosophers) is *what attention and conscious control are for, given that so much happens outside their purview*? (Dehaene 2014: 89-114).

However, even before we reach this fundamental question at the boundary between cognitive science and philosophy another issue has stimulated thoughts from both fields. How it is possible that a largely automatic creature could develop detailed non-automatic, albeit normative, decision-making models like EUDT or the belief-desire law? What is the relationship between the automatic, the heuristic, and the normative systems of decision making if they are all products of the human mind?

2.7 Dual-Process Models

One way to reconcile observations concerning heuristic judgements and automatic decision-making on the one hand with normatively constrained reasoning on the other is to divide our decision-making processes into two categories – two sets of processes, perhaps even two separate (but complementary) cognitive systems.

Kahneman (2011), for example, suggests that cognition consists of a fast, largely unconscious, economical (in terms of cognitive resources) but error prone "System 1" and a slow, deliberative, more accurate (by the standards of normative rationality) "System 2". System 1 comprises the kinds of cognitive process that humans share with many other

[64] Baddeley, A. (1996). Exploring the Central Executive. *Quarterly Journal of Experimental Psychology, 49,* 5-28.

animals. System 2 is more characteristic of – perhaps essential to – *human* reasoning. System 2 is the kind of cognition that Aristotle referred to when describing humans as "the rational animal". Classification of processes into System 1 and System 2 is also a feature of the dual-process model developed by (Evans 2008, 2009a) although, as he points out, there is some dispute at the margins as to which processes belong with which system.

Whereas a *system* may be differentiated by certain features (speed and/or resource consumption, for example), *processes* are distinguished by the set of *rules* that determine their operation. EUDT is built from a set of process rules (Bayes theorem, the sum of probabilistic utilities and the procedure outlined above). It is possible to read the belief-desire law as a process rule in itself: "do whatever you will believe will bring about your desire"[65]. Heuristics and automatic processes share the characteristic that they use non-entailing rule sets to deliver their inferences, which is one reason that they are often grouped together under System 1, as well as the fact that they are both fast and less resource-hungry (which are characteristics of that system).

Many inferences are thought to use processes drawn from both systems (Evans 2008; Kahneman 2011).

The question is whether *System 2 processes act as a normatively rational check on the more error-prone outputs of System 1*. Is System 2 *activated* (by some mechanism yet to be uncovered) whenever there is enduring uncertainty about a System 1 judgement? Or would we be better to regard System 2 as an extra resource that can be brought to bear on the decisions that System 1 delivers when **consciously directed** so to do? V. A. Thompson et al. (2011) report the results of experiments that, they contend, support the hypothesis that a "metacognitive judgment about a first, initial model determines the extent of analytic engagement." As Evans (2008) puts it:

> While some dual-process theories are concerned with parallel competing processes involving explicit and implicit knowledge systems, others are concerned with the influence of preconscious processes that contextualize and shape deliberative reasoning and decision-making.

[65] Underpinned by rules for identifying the relevant desire and for the formation of an appropriate belief.

Evans (2006) proposes that the two sets of process stand in an *interventionist* relation. Our first estimations of the problem and our *mental models* (Johnson-Laird 2006) of the prevailing situation are delivered without conscious awareness, and so are the product of System 1 processes and of *heuristic* rule sets. If time and cognitive resources are sufficient and the critical nature of the decision renders it necessary, the slower, more deliberate and normatively rational *analytic* rule set is brought into operation to check on the heuristically derived decision and to override it if necessary. Thus Evans describes his model as *heuristic-analytic:* any action-choice that is processed by the analytic system will be delivered by that system and not passed back to the heuristic system.

Conversely, other theorists (Bargh and Chartrand 1999; Saunders and Over 2009) propose that the two sets of process work in *parallel*. Every result that is delivered by an automatic or heuristic process – including intermediate results – is both available to and determines the starting position of deliberative, conscious processes. The two systems monitor one another and exchange processed information. All inferences are collaborative between the two kinds of reasoning and can be delivered - and so manifested in behaviour – as a consequence of the completion of a process in either system.

Serial and parallel families of dual-process model also agree that although decisions (and so actions) can be generated by System 1 alone, System 2 is *always* dependent on the output of System 1 to get going (Evans 2003, 2009b; Kahneman 2011; V. A. Thompson et al. 2011; Verschueren and Schaeken 2010).

Not everyone agrees that we need to posit twin processes in order to explain the apparent dichotomy of automatic and controlled processes. For example, Neumann (1984) argues that the evidence that tasks that were once only capable of being performed under conscious attention can become automatic with practice – together with the observation that there seems to be no essential difference between automatic and controlled tasks – indicate that there is *one system* and that this system is capable of being brought under conscious attention when necessary. Central control is not a separate system, but a feature of human cognition. It is, for Neumann, not possible to allocate tasks to one or the other system because unified human cognition operates over every task – with varying degrees of attention (Sykes 2006: 191-95). For a more recent but related take on the evidence, see Kruglanski and Gigerenzer (2011). Gigerenzer and Todd (1999a: 20) question the "fiction" of dual process models on the basis that they rest on a false premise:

> The unquestioned assumption behind these theories is that the more laborious, computationally expensive, and nonheuristic the strategy, the better the judgements to which it gives rise. This more-is-better ideology ignores the ecological rationality of cognitive strategies.

However this debate cashes out, it is widely accepted that while some of our complex judgements and decisions take place under conscious and deliberative control, many take place unattended (Evans 2003; Nickerson 2008). Neither has the monopoly on rationality.

Dual process models are sometimes offered as an explanation of why we can describe the normative rules and yet so frequently fail to live up to them. Speekenbrink and Shanks (2013: 682) observe "… people's preferences are often unstable and subject to various influences from the method of elicitation, decision content and goals." Or as Keren and Teigen (2004: 104) put it:

> Evidently, people are not always able to follow the prescriptions of normative theories (despite the fact that these were originally constructed by the human mind) as is assumed by standard economic theory.

Which is why the notion of **ecological rationality**, introduced in the preceding section, is so powerful. Here we have a model of judgement and decision-making that can explain our species' evolutionary success, our ability to shape the world to our needs and to adapt to almost any environmental challenge, our achievements in science, arts and philosophy all without the assumption that we are always, or even most of the time *rational* according to the requirements of our *normative models of reasoning*. Those models, which determine *what an ideal decision maker might decide under ideal circumstances* should be listed among our species' greatest intellectual achievements. Nevertheless, we should not become misled by the assumption that the ideal is **descriptive** of the way that real people decide and make judgements in real-world situations.

2.8 Chapter Summary

This chapter began with an outline of the highly influential Expected Utility theory of human judgement and decision making (section 2.1). EUDT was originally put forward as a normative model, but has become – certainly for the purposes of economic modelling – regarded as sufficiently descriptive to be used as the basis of predictions of the choices that people will make, at the aggregate level of populations if not at the individual level. I argue

that this model is very closely allied with the belief-desire law picture of action and action choice (2.2) and that both are vulnerable to similar challenges.

The first wave of challenges emerges from the intrinsic requirement of EUDT that people are adept at estimating probabilities (2.3-4). After introducing Bayes' theorem as the accepted standard for the calculation of dynamic probability, I presented several investigations in contemporary cognitive psychology that indicate that people are generally poor at estimating probability. Since the EUDT demands that the maximisation of utility can only be achieved through the accurate assignment of probabilities (both that a given choice will lead to a given outcome and that a particular outcome will deliver a precise value of utility) the prospects for that theory as a *description* of judgement and decision-making are affected by these empirical data.

These results have also motivated cognitive psychologists to develop and investigate new models of judgement and decision-making. The first of these to be considered was the *fast and frugal heuristics* programme (2.5). People use heuristic strategies of decision making:

i) Under conditions of uncertainty – where information is scarce.
ii) When under time pressure.
iii) Where cognitive resources are less than optimal – such as when individual making the choice is tired, distracted, overwhelmed by other inputs etcetera.

In short, heuristic strategies are used in the majority of real–world judgement and decision tasks. Unlike normative models (such as EUTD) heuristics imply a set of non-entailing rules that produce good enough or *ecologically rational* inferences.

Cognitive psychologists have uncovered a series of such heuristics. These provide individuals with an *adaptive toolbox* to suit many decision tasks and can also can be analysed into their process rules by the way they are used in many real-life situations.

Further scientific work on decision strategies has drawn attention to the fact that a great many of our decisions take place outside conscious awareness (2.6). Investigations into *automaticity* have shown that the processes underpinning such choices have clear identifying characteristics and are immensely powerful. Some researchers maintain that most of our cognitive processes are automatic.

Neither EUTD nor philosophical folk psychology play essential roles in the formulation of some important psychological explanations of human judgement and decision making. There are as number of predictive-explanatory accounts in contemporary cognitive psychology which have enriched our understanding of how real people, in the real world make decisions that do not feature these terms at all. Even in the case of dual-process models (2.7), few if any researchers now believe that any of our decisions are entirely the product of the slow and deliberate "system 2" – which would be where decision processes that most closely resemble the normative models would take place. Our complexity and the range of influences that play upon our reasoning ensure that normative models – even if they prescribe the *best kind of decision making* (which is doubtful in the light of work on *ecological rationality*) – do not comprehensively describe how we decide.

3 Chapter Three: Interpersonal Understanding of Action; "Attribution Theory"

Abstract: The prevalence of mind-detection and the identification of agents sets the scene for a discussion of attribution theories in social psychology, which are the ways that the science accounts for our understanding of the causes of events and, especially, actions. Two families of causal attribution with regard to action are discussed: ***personal attributions****, in which features of the agent are considered as causes and* ***situational attributions*** *in which agents are caused to act by the circumstances in which they do so. Examples of two classic investigations into the power of situations to direct behaviour are introduced along with two biases or errors of attribution: the fundamental attribution error (correspondence bias) and the actor-observer difference (illustrated with experimental investigations) which show how the attribution of causes to agents takes place in social psychology without reference to belief or desire.*

No man is an island, entire of it self.

John Donne (1623), *Meditation XVII.*

3.1 Interpersonal Understanding and the Appeal of "Agency"

Interpersonal understanding is a theory-neutral term[66] that describes an essential aspect of human existence. As social creatures, much of our lives are dependent on others in one way or another. Most people,[67] most of the time find understanding other people, including being able to predict what they will do under particular circumstances, relatively effortless and straightforward. Interpersonal understanding is an essential component in learning and teaching, the transmission of information, perhaps even language itself (D. A. Baldwin 2000).

[66] Which is why I employ it in preference to, for example, "theory of mind" or "mind-reading".

[67] Exceptions include those people who exhibit social deficits associated with the autistic spectrum.

For all that, philosophers and psychologists have long argued about the precise set of skills on which interpersonal understanding rests. Research into the *methods people use to explain action and the cognitive processes involved in this critical social skill* fall under the heading of **attribution** or **attribution theories** in **social psychology**. Försterling (2001: 17) defines these research projects as "…the scientific study of naïve theories and common-sense explanations". I contend that an implication of this definition we would expect one of its goals to be establishing an empirical foundation for philosophical folk psychology *if* belief-desire psychology accurately described these "naïve theories and common-sense explanations".

In a ground-breaking set of experiments that are regarded among the founding investigations of the discipline that was to become social psychology, Heider and Simmel (1944) examined how people describe apparent behaviour. Subjects were shown short films depicting a series of animations. In each film, a number of geometric shapes moved around the screen, sometimes interacting with each other and with a "box" shape with a single "entrance" or "exit" in ways that simulated physical reactions (collisions, obstruction and so on) and sometimes in ways that implied perceptual detection of each other (suddenly changing direction to avoid a collision, for example).

The film was shown three times to three groups of subjects. The first group were given the most general instruction; to "write down what happened in the picture". The second running of the experiment required a little more interpretive work on the part of the participants. Their instruction was to "interpret the movements of the figures as actions of persons". Participants in the main study were also asked to complete a survey after viewing the film which included the following questions:

1) What kind of a person is the big triangle?
2) What kind of a person is the little triangle?
3) What kind of a person is the circle (disc)?
4) Why did the two triangles fight?
5) Why did the circle go into the house?
6) In one part of the movie the big triangle and the circle were in the house together. What did the big triangle do then? Why?
7) What did the circle do when it was in the house with the big triangle? Why?

8) In one part of the movie the big triangle was shut up in the house and tried to get out. What did the little triangle and the circle do then?
9) Why did the big triangle break the house?
10) Tell the story of the movie in a few sentences.

(Heider and Simmel 1944: 246)

The third iteration of the experiment showed the subjects the film run in reverse and they were asked only questions 1, 2, 3 and 10 from the above list.

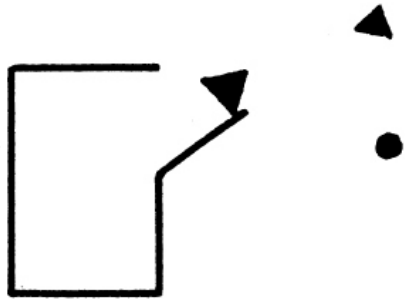

Fig. 3.1: A still from Heider & Simmel's 1944 animated film. The full film is available online at https://youtu.be/wp8ebj_yRI4.

The researchers regarded the second experiment as the main focus of their investigation, under which they wanted to know whether the subjects would interpret the film according to the "story" that they designed into it. However, the questions (1-10) are leading: it is clear from these that the experimenters expected the shapes to be interpreted as moving like intentional agents. The first experiment is more interesting with regard to how the test subjects *spontaneously* interpreted the motion of inanimate forms.

Heider and Simmel reported that of 34 subjects[68] in the first condition, "only one described the film almost entirely in geometrical terms" (Ibid.: 246). All of the others felt compelled to describe the movements of these simple shapes as if they were witnessing the interactions of intentional, agents – two of the subjects independently interpreted the movements as those of birds. Nineteen of the subjects formed their account of the action into what Heider and Simmel describe as a "connected story". There were a number of common features in these reports: the two triangles were said to have fought each other, The large triangle was described as being shut up in the house and trying to get out, the large triangle was said to

[68] In all three conditions all of the subjects were female undergraduates.

be chasing the small triangle and the circle. All of the subjects described the shapes (or "actors") as opening or closing the "door". None described the shapes as being moved by the force of the door. Several based their story on a love-triangle (in the figurative sense) in which the two triangle shapes were in competition for the affections of the circle.

The Heider and Simmel experiment has become famous (and many times replicated) because it illustrates how prevalent the **attribution of intentions** is when people come to explain movements as actions – even when, as in the case of a series of geometric shapes, it is clear that the "actors" are not minded beings.[69] This *anthropomorphism* in describing events is "…a process of inference about unobservable characteristics of a nonhuman agent, rather than descriptive reports of a nonhuman agent's observable or imagined behaviour" (Epley et al. 2007). As such, it is a manifestation of the methods that people use to explain each other's actions rather than a separate phenomenon. We are:

> *… over-enthusiastic mind-perceivers. We see minds even in objects, animals, and deities, depending on the accessibility of agency (autonomy), motivation to explain (effectance) and motivation to affiliate (sociality).* Consider everyone who talks to their plants, computers, cars, and pets as if they had human dispositions.
>
> Fiske and Taylor (2013: 152: emphasis added, references omitted)

We are prone to see agents – beings with minds and intentions – all around us. This is a demonstration of our interpersonal understanding skills at work at their most basic level. My contention is that much of this **attribution of intentions** takes place without the ascription of specific propositional attitudes – beliefs and desires – in order to explain or predict behaviour. Recent investigations in social psychology support this contention, as will be examined in this chapter.

3.2 Attribution Theory

Fiske and Taylor (2013: 149) describe the "fundamental concern" of attribution as being "how people infer causal explanations for other people's actions and mental states" and

[69] I will acknowledge the objection here that the shapes were telling a story designed by the filmmakers – carrying out their intentions, in a sense.

point out that "Causal reasoning thus recruits knowledge of other people's qualities and of situational dynamics to infer an event's causes".

The origin attribution research can be traced back to Heider (1958).[70] Heider took "commonsense psychology" as the starting point of his investigation into the phenomena, by which he understood the way that people "think about and infer meaning from what occurs around them" (Fiske and Taylor 2013: 154). Heider describes the process whereby he constructed the first model of his attribution theory by listening to the way people (the folk) describe what causes individuals to act in particular ways. He sought *to establish the rules and mechanism* by which we determine whether particular causal accounts are to be regarded as **personal**, under which some feature or features of the individual agent are taken to cause them to perform a particular action (beliefs and desires, were they to feature, would be in this category), or **situational**, whereby causal explanations rely on features of the environment or circumstances in which the action occurs.

People make attributions because causal reasoning is essential to the human need to predict the future and to control events (M. Ross and Fletcher 1985). Causal attributions are essential not only so that we can answer "why" questions of the form "why did she do that?" but also to make predictions of the form "what will she do now?" Some of these questions might never occur to us consciously. Most of the time we find other people's actions – and our own – so easy to understand that there is no need to ask, let alone to answer such questions (see 3.6, below). The mechanisms underlying these unconscious inferences – and the conscious inferences that sometimes arise – are the subject of attribution research.

Gilovich et al. (2006: 339) define attribution theory as "an umbrella term used to describe the set of theoretical accounts of how people assign causes to the events around them and the effects that people's causal assessments have". Although the principal concern here is with attribution of the causes of actions, it is presumed that similar mechanisms are in play whether causal inferences are drawn to explain the actions of identifiable agents or where events occur when no agent can be discerned. This would go some way to explaining the findings of Heider and Simmel (1944) with which this chapter opened: if we use the same techniques for explaining causes whether agents are in play or not, this would explain why *agent-like motion* is sufficient to generate inferences of agency.

[70] The same Fritz Heider involved in the Heider-Simmel experiment with which this chapter began.

Försterling further describes attribution research as being concerned with "conceptions of perceived causality" – how people account for the causes of actions and the relations of those causes to the agents that carry them out – and "the determinants of causal ascriptions" – under what circumstances to people ascribe one cause rather than another.

For philosophical folk psychology, the causes of any action *qua* action *must at least include – and in some cases must comprise – the intentional mental states of the agent, particularly their beliefs and desires*. Also, for belief-desire psychology to really be descriptive of naïve theories and common-sense explanations then the kinds of explanations that attribution researchers encounter when they investigate the strategies that people (the "folk") use *ought to reference individual beliefs and desires*.

The personal/situational distinction, and the first account of the underlying rule-sets were developed by Kelley (1973). Kelley was the first to suggest the **discounting principle**, under which *people tend to discount personal causes if the situation is sufficient to explain their behaviour*. An example would be a person fleeing a burning building. We would be unlikely to trust their post-facto assertion that their egress was motivated by a desire to give first aid to their fellow survivors: running from such a conflagration is sufficiently explained by the situational fact of the fire. Gilovich et al. (2006) give an alternative example; we are likely to give less credence to (discount), or at least treat with caution, any information offered by a prisoner who is threat of torture. The personal attribution – that the person knows or believes, and wishes to make known the information that they offer – is discounted in the face of the fact that the threat of torture alone might be sufficient to bring about their openness.

The other side of the discounting principle is the **augmentation principle**. This suggests that *extra weight is given to personal causal attributions (such as something that they know, or some feature of their character) if situational factors mitigate against their actions*. In the burning building example, if we see somebody running *into* the flames, we will be likely to look for personal explanations – they are trying to rescue people still inside, for example, or they are an off-duty fire-fighter and so their training has equipped them to enter a burning building.

Gilovich et al. (2006: 345) describe augmentation and discounting as follows:

> A person's traits are discounted as a likely cause of behaviour if the behaviour goes with the flow of the situation. In contrast, a person's traits are augmented as a likely cause if the behaviour goes against the flow of the situation.

M. Ross and Fletcher (1985: 48) lay out the two dimensions along which causal attributions are categorised: the first is the **locus** of the cause – which could be "internal" to the person or agent or "external", as with causes that originate in the situation or environment. As suggested, contemporary social psychologists tend to measure the locus dimension on a scale from personal to situational. The second dimension suggested by M. Ross and Fletcher is **stability**. Causes originating in either the personal or situational loci can be *more or less stable, meaning that either they last for a considerable length of time* (and so become encoded in the long term memory of the agent or the observer) *or they are transient and fleeting*, in which case they are not sufficiently well established to be encoded in or recalled from long-term memory and so are *held in working memory for the task of making a causal attribution*. Because working memory is limited to only a few items at a time (Baddeley et al. 2009 Ch. 3), *people have a natural tendency to seek causal explanations in long term memory and so tend to prefer stable causes*. Remembering a person's stable traits or dispositions (if the person is known to us) and attributing *them* as a cause, or simply categorising a person who we don't know under a *set of stable traits based on the activation of stereotypes* – "people like *that* always act *that* way" – requires a good deal **less effort** (exhibits greater cognitive economy) than would the deduction of specific mental states, on a moment by moment basis, from the minutiae of their behaviour (Fiske and Taylor 2013: 166-67).

Attributions involving stable personal causes, such as **trait-based** or **disposition-based** attributions play a much more significant role in the way that theorists explain causal attributions than do attributions of specific mental states – such as beliefs and desires (Fiske and Taylor 2013 Ch. 6; Gilovich et al. 2006 Ch. 9; E. R. Smith and Mackie 2007: 73-77).

We might summarise current attributional research like this:

> Much attributional reasoning is effortless and *virtually automatic*. Much attributional reasoning *focuses on inferring other people's dispositional qualities*. Much causal inference is *domain specific, not abstract and generic*. Explicit attributional reasoning is reserved for special occasions, most notably when unexpected and negative events occur. And like all social reasoning,

> attributional reasoning is *inherently social. When at a loss to explain an event, we ask someone*.
>
> (Fiske and Taylor 2013: 153, emphasis added)

At this stage I would draw attention to *one of the significant ways in which this approach to the attribution of causes diverges from the ascription of beliefs and desires as an account of interpersonal understanding*. The suggestion that causal inference is **domain specific** is in direct contrast to the idea of metaphysically essential action-causing category of beliefs and desires. There is no single set of individually necessary and collectively sufficient causes of every human action that is applied to make sense of an individual action. Our causal explanations of action depend on who we perceive to be the agent – including our relationship with and prior knowledge of them – and on the circumstances under which we perceive them to be acting.

3.3 Historic Investigations into Situational Attributions

Two of the most famous investigations in the history of research into social psychology illustrate the importance of *situational* factors – including their relationships with others – to people's choices of behaviour.

Milgram (1974) describes a series of experiments carried out in the early 1960s to investigate the phenomenon of *obedience to authority* and the suspension of individual moral judgement that often appear to accompany it. Milgram's investigation was motivated by the widespread revulsion felt within Western societies at the atrocities carried out in Europe under the fascistic regimes of the 1930s and 1940s. Milgram sought to understand how people could have become complicit in this horror and why the infamous **Nuremberg defence** – "I was only obeying orders" – should carry little or no weight. Milgram (1974: 179) expected his experiments to show that outside pressure had limits in its power to overcome ordinary moral constraints. His methods and his results have become notorious and, since concern with research ethics has become an essential part of the contemporary academy, are unlikely to be replicated.[71]

[71] The only replications I am aware of in recent years have been for "entertainment purposes" – As in the case of Derren Brown, *The Heist,* first broadcast in the UK on Channel 4 Television on 4th January 2006.

The subjects of Milgram's experiments were told that the experiments were designed to test the effect of punishment on learning outcomes: it had been advertised as a "test of memory". On arrival, subjects were selected by lot to play the role of "teacher" or "learner". This was a sham. The lottery was rigged so that the "learner" was a confederate of the researcher – the same genial, middle-aged man in each case, to eliminate the identity of the "learner" as a variable. Actual subjects were always, unwittingly, assigned the role of "teacher".

Authority was represented in the test scenario by an experimenter, again the same person in each test. This individual was dressed in a white lab-coat and carried a clipboard, conforming to the stereotype of a "scientist".

Each subject, in their role as teacher, was taken to a room and shown an impressive-looking machine with dials and a row of switches, marked from 15 to a maximum of 450 volts in 15 volt increments. The subject was strapped into a chair and given a brief electric shock of 45 volts as an illustration of the effect that the punishment would have on the learner. This was to be the only genuine electric shock administered. The "learner" was taken to the next room.[72]

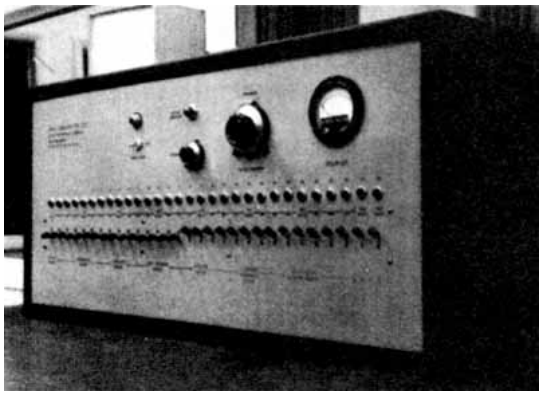

Fig 3.2 The machine operated by the "teacher" (subject) in the Milgram experiments.

The subjects were told to ask the "learner" a series of questions which required them to memorise and repeat pairs of words. Incorrect answers were to be "punished" by the teacher flicking the switch to administer an electric shock of the corresponding voltage, beginning at 75 volts. The voltage was to be increased with each subsequent incorrect answer, which

[72] In some later variations to test the effect of proximity the "learner" was in the same room as the "teacher".

had been prearranged with the complicit learner. The learner acted out a pre-scripted response to these "shocks", from a mild yelp to screams of pain and entreaties that the procedure be stopped as the "voltage" increased to around 270 volts.

Sixty per cent of the subjects (in most of the tests) continued to administer shocks all the way to the maximum of 450 volts – long after the learner had fallen silent and might, as far as the subject knew, have been unconscious or even dead. The subjects were told to interpret a non-response as an incorrect answer and so to flick the switch for the next voltage. This was one of a limited set of scripted interventions made by the white-coated experimenter during the procedure. Others included telling the subject that the shocks were "necessary for the experiment" and requesting that they "continue, please". One subject asked the experimenter whether he would take responsibility (after the "learner" had stopped responding) and was told "the responsibility is mine. Please continue." His reaction was to flick the next switch, administering a further apparent shock.

According to Blass (1999), the finding that "…ordinary individuals are much more willing to obey a legitimate authority's orders than one might have thought remains an enduring insight". This observation raises two questions. Firstly, why would we have thought that they would be less likely to obey the orders of a "legitimate authority" than appeared to be the case? This question is related to our moral expectations and, to a degree, to our *wishful thinking* that people are *more morally constrained than real-life experience would suggest*. This is outside the scope of the present discussion. More pertinent to this investigation is the second question: what factors determine the action choices that people make, even when those action choices countermand their moral considerations?

It is *possible* to present Milgram's findings in belief-desire terms. For example, the subjects *believed* that they were being directed to administer the shocks by a scientist and *desired* to please this authority figure by complying. Perhaps they further believed that no scientist would allow the experiment to get to the stage where the "learner" would suffer lasting injury. This latter point would also require them to *set aside their knowledge that giving somebody an electric shock of 450 volts is dangerous, in favour of their trust of the actor in the white coat*.

Milgram (1974: 172) investigated the role of belief in the experiments as part of his debriefing of the subjects only to the extent of confirming that the subjects believed that the

experimental setup was genuine – that they had administered real electric shocks. He found that 80% would admit to believing that they had been inflicting real punishment.

Milgram suggested two possible explanations for his findings. The first was the **theory of conformity** which suggests that, especially in stressful or crisis situations, individuals tend to *pass responsibility for decisions to their peer group or to the prevailing hierarchy* (Milgram 1974: 113-15). We become conformist, not only with the authority figure but with what we believe our peer group would expect of us – although we retain sufficient autonomy to feel that the choice to conform is our own (Milgram 1994). The second idea he offers is the **agentic state theory** (B. E. Collins and Ma 2000: 69-71; E. R. Smith and Mackie 2007: 374). This proposes that under particular circumstances, again, usually under stress but always in the presence of an identified authority figure rather than merely where social roles are defined, individuals *cease to be or even to regard themselves as autonomous intentional agents in their own right, instead becoming vehicles of the authority figure's agency*. Seeing themselves as instruments, they are no longer responsible for the consequences of their actions and so are able to bypass their inhibitory feelings of compassion or disgust (Milgram 1974: 43-48, 132-34).

Again, a plausible belief-desire rationalisation can be constructed for the agentic state explanation. In order to become such agents and to cede their individual agency to the authority figure, it could be claimed that they had to *believe* that they were no longer responsible and to *desire* that they rid themselves of blame for a morally dubious action. However, Milgram's characterisation of the agentic state expressed more in computational terms than this suggests:

> From the standpoint of cybernetic analysis, the agentic state occurs when a self-regulating entity is internally modified so as to allow its functioning within a system of hierarchical control. From a subjective standpoint, a person is in a state of agency when he defines himself in a social situation in a manner that renders him open to regulation by a person of higher status. … An element of free choice determines whether the person defines himself in this way or not, but given the presence of certain critical releasers, the propensity to do so is exceedingly strong and the shift is not freely reversible. Since the agentic state is largely a state of mind, some will say that this shift is not a *real* alteration to the state of the person. I would argue, however, that these shifts in individuals

> are precisely equivalent to those major alterations in the logic system of … automata.
>
> Milgram (1974: 133-34)

This reading sees the agentic state as *an alteration to the kinds of inferences that the individual is capable of making*. It describes a situation in which the alleged entailments of the belief-desire law break down because what the individual believes or desires are no longer relevant to their behavioural outputs. *Even if the belief-desire law was true, the **agentic state** would be a condition in which it is set aside*.

Less easy to fit into belief-desire terms is the equally infamous **Stanford prison experiment** (Haney et al. 1973). As reported in Gilovich et al. (2006: 4) this intended experiment into the social dynamics of a prison environment involved 24 male undergraduates of Stanford university, all of whom had been selected on the basis of their previous good character and screened for any psychological frailties. They were paid to play the part of a "guard" or a "prisoner" in a simulated prison. Which subjects would fill which roles was determined by the flip of a coin. Although intended to run for two weeks, the experiment had to be abandoned after six days because the behaviour of the "guards" towards their charges had begun to result in noticeable signs of stress among the "prisoners". Verbal and physical abuse had escalated to the point where the students posing as "prisoners" had been blindfolded, stripped naked and forced to simulate sex acts in scenes that, as Gilovich et al. (2006) point out, were eerily prescient of the real-life events at the Abu Ghraib prison in Iraq during the US occupation in 2004.

Both groups in the Stanford prison experiment *knew that they were engaged in a simulation*. The "guards" did not *believe* that their charges were criminals who, in some way, *deserved* the treatment that they administered. It is equally unlikely that, having signed up for the experiment, the subjects who played the guards *desired* to inflict humiliation on their peers.

The Stanford Prison Experiment is frequently held up as illustrative of the power of the situation to overcome normal moral constraints on behaviour (Carnahan and McFarland 2007; Zimbardo et al. 2000). As reported in the original paper "The environment of arbitrary custody had a great impact upon the affective states of both guards and prisoners as well as upon the interpersonal processes taking place between and within those role groups" (Haney et al. 1973).

Whatever factors were at work in these infamous and probably unreplicable investigations, it is difficult to make a case that the essential role of beliefs and desires described by philosophical folk psychology is up to this task. Experiments like Milgram's and the Stanford Prison have provided the impetus to new directions in research into how we attribute causes to actors.

3.4 Fundamental Attribution Error or Correspondence Bias

The realisation that situations have such a marked effect in the actions that actors choose, led researchers to investigate *why it is that we seem to prefer personal attributions*. After all, one of the anticipated outcomes of the Milgram experiments was that there would prove to be some fundamental difference in the way that people from one culture (Germany in the 1930s) would respond to authority, when compared to another (Americans in the 1960s). In the process, social psychologists uncovered a fascinating bias in our attribution strategies. The **Fundamental Attribution Error** (L. Ross 1977) or **correspondence bias** (Gilovich et al. 2006: 356) describe an inferential bias to which our attribution strategies are prone simply because of the mechanisms that underlie them. It suggests that *people have a preference for causal explanations of an action based on features of the agent (personal attributions) even in circumstances where the situation and an agent's reaction to it offer sufficient explanation for what happened*.

Superficially the correspondence bias is reminiscent of the way that propositional attitude ascriptions are generated by the belief-desire law according to philosophical folk psychology. The assumptions underlying the FP picture that generate these kinds of inference are summarised in this quotation from Horgan and Woodward (1991: 149 emphasis added):

> Whatever else a person is, he is supposed to be a rational (at least largely rational) agent – that is, *a creature whose behaviour is systematically caused by, and explainable in terms of, his beliefs, desires, and related propositional attitudes*.

If this is true, then the observer of an intentional action would be compelled to seek out beliefs, desires "and related propositional attitudes" that make sense of the action – that "fit" according to the schema. We would maintain that those particular beliefs and desires etc. were present even if the agent denies holding them and even if their behaviour is fully

accounted for by the situation in which they acted. The person did not run from the building because it was on fire (situational attribution) – an action which is entirely rational in terms of survival. They left the building because they *believed* that it was on fire and *desired* to avoid being burned and *believed* that leaving hurriedly was the best way to avoid being burned (unstable personal attributions); even if their best recollection was of seeing the flames, feeling the heat and running in a state of panic. This is almost the same effect as the tendency to make stable personal causal attributions even when situational causes are sufficient to explain the action. The only difference is that *it is much more easy to refute an erroneous stable personal attribution than it is to cancel out a set of propositional attitudes that, as well as transient and unstable, might even occur beneath the subject's conscious awareness*.

Two reasons have been suggested for the Fundamental Attribution Error/Correspondence bias (Gilovich et al. 2006: 360-66): The first is the so-called **just-world hypothesis**. This suggests that many of us have an in-built tendency to regard what happens to anyone as, in some way *deserved*. An insidious and unpleasant effect which has been attributed to the just world hypothesis has been the tendency of many commentators (and, sadly, of some people involved in the criminal justice system) to express the view that victims of rape must bear some responsibility for their having been attacked (Furnham 2003). Belief in a just world would mean that even when people are entirely at the mercy of situational causes, *they must possess some feature, as individuals, that has brought about the result*.

A second reason for the fundamental attribution error is the higher **perceptual salience** of persons compared to situations: "People are compelling stimuli of considerable potential importance to us" (Gilovich et al. 2006: 366). We can draw impressions of people and their behaviour directly through perception, often with limited need to infer unseen elements. Situational facts are not so readily presented to perception and so we are reluctant to place them at the fore when seeking explanations.

We are also reluctant to correct biased attributions. Doing so is effortful, time consuming and hungry for cognitive resources which might not be available to replace personal (dispositional) attributions with situational attributions, even when the latter might be more successful as an explanation for the action (Geeraert et al. 2004).

Van Boven et al. (1999) set out to show that although individuals might be prone to the correspondence bias, they would be unaware of its effects. Accordingly, in a scenario where

situational factors clearly played the dominant role in their own actions, it was expected that they would discount the effects of the correspondence bias when predicting the attributions made by those who observed them. Their findings appear to contradict this expectation.

In their first study, involving a total of 92 participants, a group of twenty were pre-selected to play the role of "speakers" according to their attitudes to **affirmative action** – the controversial policy of giving preferential treatment in recruitment or job promotion to female candidates or those from minority groups in order to redress imbalances which have arisen from many years of discrimination. All of the candidates had been pre-selected by questionnaire (at a time unrelated to the study so that the connection was unknown to them), according to the extreme nature of their views on the subject – either for or against.

Each speaker was asked to write and to deliver a videotaped speech either supporting or opposing affirmative action for academic admissions. Without the speakers' knowledge, the experimenters ensured that those speaking in support of affirmative action were those identified as the most vehemently opposed to the policy and those speaking against the idea were those who had previously expressed the strongest support. Each speaker was given thirty minutes to write their own address, based on a model provided by the researchers. They were encouraged to incorporate as much of their own thinking and ideas on the subject as they could (while remaining within the pro- or anti- brief).

An experimenter introduced each speaker on his or her video by saying:

> "This is speaker number [x]. We have asked speaker number [x] to write a short speech entitled 'why colleges and universities should [should not] use affirmative action policies in their admissions policies."

It was made clear to the participants who would watch the speeches on video that *the speakers were not expressing their own opinions but presenting the view that they had been given*. They were, in other words, *operating under situational constraints*.

After watching each video, the participants were asked to rate each speaker against the statement "Speaker [x] is a supporter of affirmative action laws for hiring women and minority individuals" on a thirteen-point scale from "The speaker doesn't agree at all" (-6) to "The speaker agrees very much" (+6). The speakers, after making their recordings, were told that other students would view their speeches and attempt to discern their true opinions.

Each speaker was asked to predict, against the same scales, how the observers would rate their true attitude to the subject.

Initially the experimenters had expected the speakers not to anticipate the extent of the correspondence bias, given that they knew that the observers were fully aware of their situational constraints. In fact, not only did the results show a clear correspondence bias on the part of the observers (a tendency to attribute dispositions to the speakers which corresponded to the content of their speeches), the *speakers* **overestimated** the extent of that bias *by a factor of more than two*. This in spite of the fact that the speakers had been made aware that the observers *knew* that they had not chosen the position they expressed (their situational constraints).

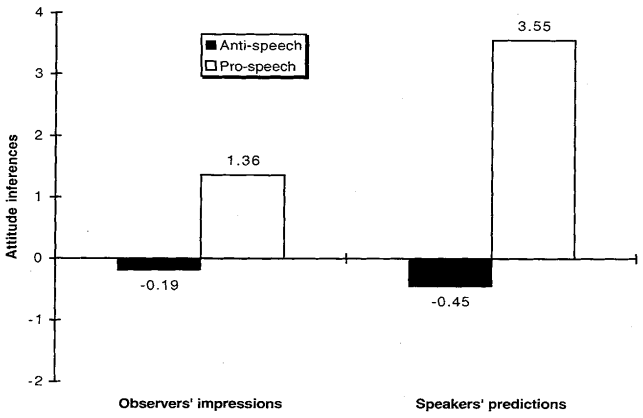

Figure 1. Observers' inferences of speakers' attitudes toward affirmative action and speakers' predictions of observers' inferences.

Fig. 3.3 Results of first study from Van Boven et al. (1999: 1190)

The second study reported in Van Boven et al. (1999) was based on an approach from Gilbert and Jones (1986).

Participants were divided into pairs (screened to ensure that they did not know one another) and the members of each pair were randomly assigned the roles of "questioner" and "responder". Communicating via an intercom system, the questioner asked the responder a series of twenty questions about their general life attitudes, including some indicating their moral outlook, such as:

> Do you consider yourself to be sensitive to other people's feelings?

Each responder was equipped with a set of the questions and a set of scripted answers. Two possible responses were available for each question, one of which was designed to be

altruistic and caring, the other *selfish and cold-hearted*. For the example given above, the two possible answers were:

> I try to be sensitive to other's feelings all the time. I know it is important to have people that one can turn to for sympathy and understanding. I try to be that person whenever possible.

Or alternatively:

> I think there are too many sensitive, 'touchy-feely' people in the world already. I see no point in trying to be understanding of another if there is nothing in it for me.

This would be a straightforward "question-response" type test except for this twist; The *questioners* were equipped with a signal light system, *whereby they could tell the responder which kind of response to give*. In half of the cases, the responder was told by the experimenter to signal for 80% selfish responses, in others for 80% altruistic responses.

After each set of 20 questions, the questioners were asked to assess the responders' "true, underlying character" against a series of traits – likeable, trustworthy, selfish, greedy, dependable, altruistic and kind hearted – on a 13-point scale for each measured from "not very" (0) to "very" (12). Each responder was given the same set of measures and asked to predict how the questioner would rate them.

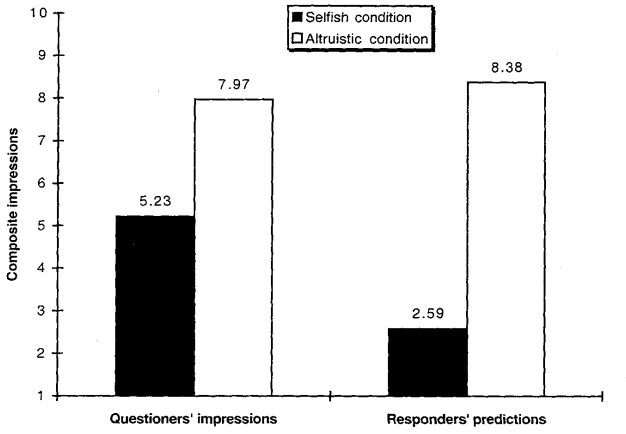

Figure 2. Questioners' impressions of responders and responders' predictions of questioners' impressions.

Fig 3.4 Results of second study from Van Boven et al. (1999: 1192).

The participants on both sides knew that the kinds of response that were given were entirely determined by the situation; indeed, the *questioners* were responsible themselves for

determining whether the answer to each question would be "altruistic" or "selfish". And yet *still* they tended to think more highly of the real character of those who had read out predominantly altruistic responses. The participants seem incapable, even in this extreme example, of denying that something about the individual determines their responses. If you give mostly altruistic replies, then you are a nicer person than someone who gives predominantly selfish replies - even if I am the one telling you how to respond to each question!

The responders' predictions of the size of the correspondence bias was consistent with the first study, in that they consistently overestimated the degree of bias.

In their general observations and conclusion, the authors firstly noted that people expect the correspondence bias – that the attributions will tend to be dispositional, even when the situational cause of the behaviour is explicit. They go on to suggest that, under constrained experimental situations where the speaker's or responder's self-image is at odds with (and so obscured by) their overt behaviour, they *overestimate* the degree to which the observer or questioner, blinded by the deliberate obfuscation of the "true self" would rely on behavioural cues. These findings were at odds with their *original hypothesis that people would underestimate other people's susceptibility to the confirmation bias*.

Explaining these findings in belief-desire terms is, I suggest, a considerable challenge. We would have to presume that observers and questioners in all of these studies based their assessments of the speakers' and responders' characters on an assumption that in order to speak the words they must, on some level, *believe* them. And what of desire? Are we to understand that speakers desire to be believed by their audience so much (even when they are uttering words utterly foreign to their own attitudes) that they predict that people will attribute their behaviour to their dispositions to an even greater degree than they actually do?

> We consistently found that people not only anticipate the correspondence bias, they tend to overestimate its magnitude. We attributed this overestimation to people's oversimplified intuitive theories that others are inveterate dispositionalists who give little regard to situational influences; we obtained support for this contention by showing that overestimation of the correspondence bias does not stem from actors' evaluative concerns.
>
> (Van Boven et al. 1999: 1198)

Intriguingly for those who regard belief-desire reasoning as an innate feature of human social cognition there were marked differences in the scale of the estimates for the correspondence bias when similar (translated) studies were carried out with Japanese students: They report that:

> Japanese participants, whose folk theories about the causes of behaviour emphasize situational factors more than Western folk theories, did not overestimate the magnitude of the correspondence bias.
>
> (ibid.)

This evidence of cultural difference in attribution strategy supports the contention that I will make more overt in Chapter Four that far from reflecting an essential feature of human social cognition, the idea of a folk-psychology based on the ascription of beliefs and desires is a learned **cultural artefact**: *part of the stock of stories and story archetypes that make a significant contribution to culture.*

3.5 Actor-Observer Differences

Another observed bias in our attribution strategies presents a further challenge to the assumptions of belief-desire psychology. One of the appeals of the belief-desire model of intentional agency is its inherent symmetry. It is presumed that *if* on the basis of our introspective experience of our own intentional actions we notice that when we desire something we tend to do whatever we believe will satisfy that desire, *then* we are justified in assuming that the same goes for other agents whose actions we witness. Because we introspect the causes of our own actions in this way, we can extrapolate that any agent's desire for something, guided by the belief that a particular action will fulfil that desire, will cause them to act.

Unfortunately for this picture, research in social psychology has uncovered considerable asymmetry between the causal attributions that people make to explain their own actions and those that they use to account for actions that they witness. This effect is known as the **actor-observer difference** (Gilovich et al. 2006: 367; E. R. Smith and Mackie 2007: 101). At its simplest, this observation suggests that *people are much more likely to attribute their own actions to situational factors while tending to attribute the actions that they observe in others as being caused by personal traits*. Jones and Nisbett (1972: 80) argued from their experimental investigations that "…there is a pervasive tendency for actors to attribute their actions to situational requirements, whereas observers tend to attribute the same actions to stable personal dispositions." If the belief-desire picture were even half right, in that we begin to understand the causes of action by introspection, then we would, I suggest, expect this pattern to be reversed. Also, according to philosophical folk psychology, the belief-desire law applies equally to explanations of our own actions (reason-giving) as to our interpersonal understanding of the actions of others. The asymmetry of the actor-observer difference is a direct challenge to this picture.

For example, Nisbett et al. (1973) asked a group of U.S. male undergraduates to explain their own choice of college major and their choice of romantic partner[73] and to describe what had motivated their best friends' choices in the same areas. Even in the case of people they knew well, the subject attributed their own choice of partner to twice as many features of that person than to their own "dispositions" while in the case of their friends they offered a similar number of reasons for each. In the case of their choice of college major, *the subjects gave three times as many features of the course as reasons for their choice than they thought might explain their friends' choices*.

To revisit the burning building example, when someone downplays their "heroic" action in saving others with words like "I'm not a hero: I just did what anyone would have done" they might not be exhibiting false modesty. When they go on to say that "the fire-fighters who went into the building after I got out, now they are the real heroes" this might be a manifestation of the actor-observer difference. Likewise, if we ask the "heroic" fire-fighters why they went into a burning building they might say "the building was on fire and it's my

[73] This being the 1970s the word used in the study was "girlfriend".

job" – again preferring situational rather than personal attributions in accounting for their own actions.

Gilovich et al. (2006: 367-69) posit four factors in concert that play a role in generating the actor-observer difference:

i) actors and observers might have different understandings of what they need to explain. Actors might take their stable dispositions (traits) for granted and so not necessary to any explanation. The fire-fighter who says "that's my job" doesn't need to add that they are also *the kind of person* who becomes a fire-fighter in the first place.

ii) Actors and observers might have different *perceptual saliences*. The actor is attending to the situation, its opportunities to act and the constraints it places on their possible courses of action. They do not see themselves in the setting. Observers, on the other hand, tend to focus on what the actor is doing. They are better placed to identify not only what the actor does, but what personal attributes might be at play.

iii) Actors and observers have different information about the action. Actors might consider that they have access to their intentions (although, as we will see below, this is unlikely to be infallible) and so give greater weight to the situational factors that either motivated or thwarted their intentions.[74]

iv) The false consensus effect (Gilovich et al. 2006: 368) is another well-documented social cognitive bias through which individuals tend to regard their own actions as more "typical" than those of others.[75] When an observer witnesses an action that is at variance with how they (or any of their friends) might expect to act in a similar situation, this "low-consensus" behaviour is more likely to be attributed to dispositional factors than to the situation.

The role of **saving face** (Försterling 2001: 87-91) should also be considered: this is thought to be an important contributor to causal attribution strategies (Brockner et al. 1981; E. R. Smith and Mackie 2007: 131). If there is a risk that one's action might be judged negatively,

[74] Which suggests that we might seek to attribute the outcome to situational factors even when taking unstable personal factors (mental states) into account.

[75] Because we tend to socialise with people who share our attitudes, this sense that we are just like "everybody else" is reinforced by our peer group.

it is more face-saving to blame the situation than to accept "that's what I am like" – as would be implied by a trait attribution. In the burning building example, an individual who fled the building alone without trying to help or to lead others to safety might claim that "the heat was so intense and there was a real danger of collapse" (situational factors) while anyone who saw his exit might call that same individual a coward (dispositional attribution). Face-saving might help to explain Malle (2006), who found, through a meta-analysis of 173 actor-observer studies, that *the bias is most marked where actors might be susceptible to unflattering dispositional attributions*.

Charged with researching issues of road safety and, in particular, the contribution that driver aggression makes to road traffic accidents, Lennon et al. (2011) started from a hypothesis that differing accounts of the causes of accidents (due to actor-observer differences) might explain why aggressive drivers seem not to learn their lesson even when their attitude has resulted in an accident.

Their method was to invite 193 drivers to take part in a study. The subjects were divided at random into two groups, with the members of one group each being assigned the role of "instigator" and those in the other group given the role of "recipient". The two groups were given eight descriptions of driving scenarios that had been identified (by previous research) as examples of aggressive driving. Recipients and instigators were given versions of the scenario worded appropriately for their role. For example:

> Scenario 8 (recipient perspective); You are in the left lane behind another vehicle. When the left turn arrow light is given, the vehicle does not move because the driver is not paying attention. You tap on the horn to get his/her attention and he/ she gives you the middle finger in their rear-view mirror
>
> Scenario 8 (instigator perspective): You are in the left lane waiting for a green arrow. When the left turn arrow light is given, you do not move because you are not paying attention. The driver behind you taps his/her horn to get your attention and you give him/her the middle finger in your rear-view mirror

After reading each scenario, the recipients were asked the following question:

> Thinking about the actions of the other car driver in the situation, which of the following descriptions would best explain *their* behaviour?

While the instigators were asked:

> Thinking about your actions in relation to the other car driver in the above situation, which of the following descriptions would best explain *your* behaviour?

Both groups were given four possible explanations from which to choose:

a) Bad luck.
b) The road or traffic conditions or The road sign and road markings (depending on scenario).
c) A mistake of your/their judgement at the time.
d) Shortcomings in your/their driving ability.

These options correspond to the four possible dimensions of a causal attribution. The reason given at a**)** is an **unstable situational attribution**, **b)** is a **stable situational attribution**, **c)** refers to **unstable personal traits or dispositions** and **d)** to **stable personal traits or dispositions** (Lennon et al. 2011: 213).

The conclusion of this study supported the starting hypothesis. Recipients were significantly more likely than instigators to attribute aggressive driving behaviour to poor driving skills. Instigators preferred to account for their performance in terms of the situation – explanation (b) – or, where the scenario made these unavailable, in terms of personal but unstable causes such as simple errors of lapses of judgement. If they just made a simple mistake, the researchers reasoned, why would an aggressive driver see the need to get help to modify their behaviour? What you identify as the *cause* of aggressive driving seems to depend on your standpoint.

3.6 Social Heuristics and Automaticity

Researchers have applied the fast and frugal heuristics programme, introduced in the last chapter, to social navigation and specifically to attribution strategies. In part this has been motivated by a desire to **functionally analyse** the fundamental attribution error/correspondence bias and actor-observer differences into their underlying capacities and processes, in keeping with the objective of psychology established in Chapter 1, whereby genuine causal explanations are sought for persistent effects.

Dealing with other people is a typical case of operating with less than optimal information, with limited time to make a decision and with restricted computational resources. For example, we do not always *know* what other people expect from our exchanges or have a complete history of their life prior to the interaction. Many social interactions are fleeting and we are expected to respond to people without much time to deliberate. We are also frequently preoccupied with our own goals for a particular interaction, rather than having a great deal of resources available for "reading" the other party.

Experimental results suggest that much of our interpersonal understanding depends on the application of simple heuristics rather than on the ascription of mental states. In his review, Marsh (2002) categorises the kinds of heuristic used in social settings under three headings:

i) **Search heuristics**, which are shortcuts to find meaningful knowledge is specific circumstances: an example would be to find a suitable stereotype under which to match a person to their actions.
ii) **Assessment Heuristics** are used to rank those options identified by the search according to preference or suitability to the situation.
iii) **Selection heuristics** are used to choose from a limited set of alternatives such as the choice between personal and situational causes in attribution.

Slovic et al. (2002) argue that we should not overlook the power of what they dub the "Affect Heuristic" in managing social interactions. We very quickly – on the basis of minimal cues – judge individuals we meet and even their trivial actions as "good" or "bad" on the basis of our emotional responses to them and to their actions. This colours our causal attributions to a degree that should not be underestimated. Messick (1999) suggests that decisions in social settings depend on "alternative logics" in which our perceptions of appropriateness, identity (including stereotype activation) and some socially encoded rules play decisive roles. Similarly, Garcia-Retamero et al. (2009) have investigated how, despite the "notoriously complex" nature of social contexts, we make decisions quickly and on the basis of minimal information by employing deceptively simple strategies which employ implicit (and culturally pre-determined) knowledge of how social environments are structured.

Divergences from normative ideals and other features emerge from the non-entailing nature of heuristic and implicit decision making. Some theorists, most notably Tversky and Kahneman (1974) would characterise these divergences as *errors or biases*. Most of the

time, however, these heuristics help us to avoid errors or embarrassment. They are also key to understanding each other's actions.

Fiske and Taylor (2013) dedicate a chapter (Ch. 7) to "heuristics and shortcuts" in the context of social inferences and decision making. Among the heuristics they consider are the **representativeness heuristic***,* and the **simulation heuristic**. Since each has a bearing on how we make judgements about people and their potential actions/choices, we should consider the application of representativeness and simulation to attribution theory.

The representativeness heuristic (Kahneman and Tversky 1972; Tversky and Kahneman 1974) is key to the way that people activate stereotypes. A useful illustration of representativeness is still that offered by Kahneman and Tversky. Subjects were invited to read a short description of an imaginary person:

> Steve is very shy and withdrawn, invariably helpful, but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.
>
> (Tversky and Kahneman 1974: 1124)

The subjects were then asked to choose which profession they found it most *likely*[76] that "Steve" would work in from this list: farmer, trapeze artist, librarian, salvage diver or surgeon. Fiske and Taylor (2013: 179) point out that "With adequate information about the frequency and personal characteristics of people in these occupations, one could conceivably tally up the probability of a meek surgeon, a shy trapeze artist, and so on, and calculate the odds that Steve is in each occupation." In practice this is not what we do – and not only because, as we saw in Chapter 2, we tend to be rather poor at dealing with probabilities. We cannot spare the time and effort involved in this calculation and so we ask "what kind of person do we typically expect to find in these occupations?" and then "what kind of person is Steve most like?". We then (as did most subjects in Tversky and Kahneman's experiments) answer "Steve is a librarian". Despite this being a divergence from normative rationality, we are likely, in such a case, to be right: stereotypes do not become stereotypes without some long-standing empirical justification, even if, in the case of negative stereotypes, the evidence has been exaggerated.

[76] A probability judgement.

Even though representativeness heuristic might result in persistent divergences from normative standards (Teigen 2004; Tversky and Kahneman 1974), they have considerable utility in making attributional judgements. Faced with a burning building and the background information "Steve is a librarian" we would not expect Steve to run into the flames. If, however, we know that "Lisa is a firefighter" then we would not be surprised to witness Lisa trying to enter the building to effect a rescue.

Even in less dramatic settings we frequently make *fast and frugal* trait attributions on the basis of stereotypes. Why didn't Steve go to the party? because he's shy. *Typical librarian.*

> The representativeness heuristic is a quick, though occasionally fallible, method of estimating probability via judgements of relevancy. It is also perhaps our most basic heuristic. Identifying people as members of categories or assigning meaning to actions is fundamental to all social inference.
>
> (Fiske and Taylor 2013: 181)

More negatively, representativeness is the source of prejudicial judgements on the basis of the way somebody is dressed, the way that they speak and, perhaps most sadly of all, their ethnicity. It allows us to explain someone's action with the minimum of effort.

The **simulation heuristic**[77] (Kahneman et al. 1982) is a problem-solving strategy involving the use of imagination to consider a hypothetical situation and so imagine what might happen. If I want to know how a friend is going to react if I tell her that I can't make it to her wedding, I am likely to imagine my friend and her stable traits and dispositions, and to imagine the conversation I will have with her when I tell her that I won't be there. I could even try out various excuses in this simulation and imagine her reaction to each of them before deciding whether to make up an excuse or tell her the *real* reason. According to Fiske and Taylor (2013: 184) "The simulation heuristic addresses a variety of tasks, including prediction (Will Joan like Tom?) and causality (Is the dog or the kid to blame for the mess on the floor?)." Part of the simulation is likely to be our own and the imagined participant's emotional responses to the hypothetical scenario, which allows us to give weight to the various options and to choose that which will cause the least upset for ourselves and others (for example).

[77] Not to be confused with "simulation theories" of philosophical folk psychology.

We could also use the simulation technique to imagine (rather than calculate) which course of action will entail the least effort. If we are applying this to our own action choices, the simulation will guide us to the preferred option, knowing what we do about our own preferences (traits). When it comes to predicting the actions of others, we will predict what someone known to us will choose based on their stable dispositions and the imagined or simulated features of the task: *Jenny will run to the coffee shop because she is an exercise fanatic*. *Kelvin will take the bus because he can't bear cold weather*. In the case of somebody we don't know well, we might construct a simulated scenario from the traits we have applied to them thanks to stereotype activation (through the representativeness heuristic, for example): *it's no use following the young man in the Metallica T-shirt because I doubt that he's going to the Bach organ recital*.

Much of our social reasoning also seems to take place outside conscious awareness or conscious control.

> When one considers that this automatic perception of another person's behaviour introduces the idea of action – but from the outside environment rather than from internal, intentionally directed thought – a direct and automatic route is provided from the external environment to action tendencies, via perception. The idea that *social perception is a largely automated psychological phenomenon is now widely accepted*.
>
> (Bargh and Chartrand 1999: 465 emphasis added)

Bearing in mind the features of automatic processes (**effortless, unconscious, non-intentional** and **autonomous**) introduced in Chapter 2, section 2.6, it is apparent that a good deal of our interactions with other people are mediated by these non-controlled processes, even if not all and even if one accepts that these are mediated (in part) by belief and desire ascriptions (Apperley and Butterfill 2009; Steuber 2012). Gilovich et al. (2006: 19) point out that our emotional reactions to people and to situations occur largely outside conscious control and are an essential part of our interpersonal understanding. They also suggest that some of the persistent errors and biases in attributions – including the two examined in the present chapter – occur because of a mismatch between automatic and controlled processes.

> The concepts of automaticity and preconscious processing of information help us to understand why we are often blind to the role of many important situational factors and why the processes underlying construal may be hidden from us.
>
> (Gilovich et al. 2006: 22)

The question that arises from this observation is: in what circumstances are automatic or heuristic causal attributions not up to the job? To put it another way, what circumstances *might* trigger the ascription of beliefs and desires – among other varieties of deductive reasoning – as a way of explaining the causes of a particular agent's actions? Fiske and Taylor (2013: 153) suggest that *we tend to deploy normative models as a potential explanation when there are anomalous behaviours to be explained*. For example, when we see someone run *into* an obviously burning building our conscious attention is drawn to the event and it is then that we ask "why" in a more conscious and attentive way.

It might be that the problem with causal attributions in action explanation is not so much a problem of other minds, but problems with accessing the content of our own minds with any certainty. Evidence suggests that the introspection of mental states to which we presume ourselves to have privileged access is a far from straightforward matter. As Wilson (2002: 93) puts it: "There are cases in the psychological literature of people who are so ignorant of why they respond the way they do that they have to invent an explanation."

Nisbett and Wilson (1977) contend that "…there may be little or no direct introspective access to higher order cognitive processes." Based on a review of the experimental data in social psychology to that date, this claim breaks down into three parts. Firstly, that people are not usually able to predict the effect that any given stimulus might have on their rational inferences. They might be unaware of a stimulus, of their responses to the stimulus or even that, as the authors put it "an inferential process of any kind has occurred." Secondly, they might mistake other, plausible explanatory inferences for those which were the real causes of the behaviour. They might, in such cases, be answering a different question altogether: asking "what might plausibly justify my having taken this action" is not the same as (nor likely to elicit the same answer as) "what caused me to take this action". Finally, although Nisbett and Wilson accept that "…subjective reports about higher mental processes are *sometimes* correct" this success is not, they argue, due to privileged introspective access to their mental states (perception of the mental), but an instance of an "…incidentally correct

application of *a priori* causal theories". In other words, sometimes, perhaps, belief-desire inferences get lucky!

The idea that the introspection of mental states is an inaccurate, perhaps even misleading source of data when attributing the causes of actions to oneself persists. As Gilovich et al. (2006: 21) write:

> For many cognitive processes, it seems, we cannot accurately describe what is going on in our heads. … This applies to our guesses about other people, to our understanding of how we go about making causal attributions for physical and social events, and even how we come to choose one applicant versus another for a job (or one romantic partner over another). Often we cannot even consciously identify some of the crucially important factors that have affected our beliefs and behaviour.

Heider (1958 Ch. 1) began his investigation with an analysis of common sense psychology. If belief-desire psychology does not feature in social psychology's models of attribution, that is probably because researchers have found better explanations for the target phenomena. For example:

> In the 1970s the naïve scientist view identified complex reasoning to underlie causal inference. These analyses created the impression that explicit causal reasoning is time consuming, ubiquitous and central to other inferential processes and behaviour. However, the idea that people use much of their cognitive capacity much of the time for causal reasoning is unlikely to be true. … cognitive capacity is costly … By contrast *long-term memory is virtually limitless so probably we solve many causal dilemmas simply by accessing long-term memory for causes relating to specific people, situations or events*.
>
> (Fiske and Taylor 2013: 150, emphasis added)

All of this suggests that self-reports of mental states which precede action and so might be offered as reasons or causes of that action should not be accepted as straightforwardly correct. There is no "privileged access". The Fundamental Attribution Error/Correspondence Bias indicates that we have a marked preference to make personal attributions, even when situational factors are sufficient to explain the action. Actor-observer differences point out that we often arrive at different causal attributions when we

explain the actions of others to when we explain our own. If we are also wrong, in many cases about our own motivations then the explanatory value of belief-desire psychology is in question.

3.7 Chapter Summary

This chapter considered how social psychologists investigate our understanding of our own and of others' actions – the field usually called *attribution theory* (section 3.2). Investigations into causal attribution have suggested that we make attributions of what causes a person to act on the basis of two kinds of factor: *personal* and *situational*. These two kinds are further divided into *stable* and *unstable* factors. *Stable personal factors* include individual traits or dispositions while *unstable personal factors* comprise mistakes as well as transient mental states. *Stable situational factors* include most of the situations of the environment in which the action was performed whereas *unstable situational factors* include variables such as the weather or simple "bad luck".

As shown in two highly influential experiments in social psychology (3.3), features of the social environment – situational causes, in other words – have been shown to have marked effects on the decisions that people make. Under certain conditions the situation causes them to override their own stable personality traits, to act markedly "out of character".

Two persistent errors or biases of attribution have been observed and used by attribution researchers to uncover the computational processes of attribution. The fundamental attribution error (or correspondence bias) (3.4) is a *tendency mistakenly to attribute a cause based on an agent's traits or dispositions rather than the situation under which the action takes place* or when *an inference about the traits or dispositions the person is mistakenly drawn in order that this corresponds with the behaviour to be explained*. The attribution of beliefs and desires that *fit* a behaviour – even when the agent is consciously unaware of possessing those states, is functionally similar to the correspondence bias (except the personal qualities being attributed are unstable propositional attitudes).

The second feature of ordinary causal attributions is the asymmetry between self-attributions made when the person making the attribution is the actor and those attributions that people tend to make when observing the actions of others. This actor-observer difference (3.5) is again at odds with the belief-desire picture, the inherent symmetry of which is part of its appeal.

Like the judgements and decisions examined in Chapter 2, many inferences of causal attribution are facilitated by social heuristics and many occur automatically, outside conscious control (3.6). Social psychologists have also investigated the presumed infallibility of introspection, that we are expected, mostly, to know the contents of our own minds. This is a source of the expected symmetry between reasons and genuine motivations which is foundational to the truth of the belief-desire law: when we give reason-explanations featuring propositional attitudes we are assumed to be accurately and truthfully reporting the mental states that motivated our actions.[78] Social psychologists doubt this, as well. We seem to know much less about our own motivations than was formerly thought.

[78] Even if not committed to the view that those mental states are the cause of their action.

Part Two:

Everyday

“Belief” and “Desire”

4 Chapter Four:
Narratives of Action;
Stories and our Picture of the Mind

Abstract: In this chapter, I develop the idea that the proper place of action accounts and reason-giving couched in belief-desire terms is as a species of ***narrative discourse****. As such, they do* ***not*** *serve as causal explanations but merely offer a* ***story*** *about what has happened, the acceptability of which is judged by a different standard than are causal explanations (an idea which is developed in the course of the chapter). In support, I call on the importance of narrative to the way that people construct their understanding of the world. Narratives are ubiquitous. More than this, they seem to be an essential part of the way that humans make sense of events, relationships between events and actors, including themselves and, especially, the understanding and handling of time. As an illustration of the power of narratives, and by way of contrast with the causal-explanatory commitments of cognitive and social psychology, I introduce* ***narrative psychology.*** *From narrative psychology, a number of therapeutic approaches have been developed with narratives at their heart – none of which, I will suggest, depend on fixing the metaphysical status of beliefs and desires as causes of action. I will also suggest that judging the acceptability of narratives has features in common with* ***heuristic judgement*** *and I postulate the existence of a* ***narrative heuristic****. In conclusion, the acceptability of a* ***narrative of action*** *– even one featuring beliefs and desires – does not depend on its causal-explanatory adequacy.*

The problem of how to make all this wisdom understandable, transmissible, persuasive, enforceable - in a word, how to make it stick - was faced and a solution found. Storytelling was the solution - storytelling is something brains do, naturally and implicitly. Implicit storytelling has created our selves, and it should be no surprise that it pervades the entire fabric of human societies and cultures.

(Damasio 2010: 293)

4.1 Why Narratives? Heider and Simmel revisited.

My contention in this chapter is that narratives are among the most powerful linguistic forms that we encounter – and as I will argue in the next section, we seem to encounter them all the time. Of most interest to the present thesis are what I call narratives of action – stories about what people have done and why: in particular, those that advert to the mental states (propositional attitudes) of individual agents. How are these assessed? What are our criteria for assessing them? As we shall see, the narratologist Marie-Laure Ryan suggests that it is definitive of a narrative that "the sequence of events must form a unified causal chain…" (Ryan 2007). I raise the question of whether this entails that beliefs and desires are acceptable as events in a narrative sequence (of action) in virtue of their *presumed causal role*. And even if that is the case, is the presumption warranted other than being the default within a particular cultural context? Is a statement such as "I wanted [desired] to buy milk. I remembered [believed] that the corner shop was open and so went there," a satisfactory narrative? These are the questions that this chapter seeks to answer.

Along the way I intend to establish that:

a) Narratives play a crucial role in people's relationship with the world.
b) Narratives have a notable effect on how people impose meaning on experience.
c) Narratives are mutable. We can shape our conception of the meaning of an event by making changes to the narratives that we use to describe them.
d) Judging people as minded beings and autonomous agents does not depend on the possession of a mental state vocabulary.
e) The content and form of narratives that we habitually construct and judge acceptable is culturally determined, dependent on the kinds of narratives that we have encountered since our formative years.

The discussion of attribution in chapter 3 opened with an outline of experiments carried out by Heider and Simmel (1944), to illustrate the point that we can't help explaining motion (of certain kinds) in terms of the intentions of agents. At the opening of this chapter, I want to revisit those experiments to draw attention to another feature of their findings.

The experiments involved subjects watching the apparent motion of a few geometric shapes in an animated film. In the first experiment, 36 participants were asked to describe the action of the film in their own words – to "write down what happened in the picture" (Ibid. 245).

The following description is presented in the original paper as "representative of the interpretation commonly made in the group":

> A man has planned to meet a girl and the girl comes along with another man. The first man tells the second to go; the second tells the first, and he shakes his head. Then the two men have a fight, and the girl starts to go into the room to get out of the way and hesitates and finally goes in. She apparently does not want to be with the first man. The first man follows her into the room after having left the second in a rather weakened condition leaning on the wall outside the room. Man number one, after being rather silent for a while, makes several approaches at her; but she gets to the corner across from the door, just as man number two is trying to open it. He evidently got banged around and is still weak from his efforts to open the door. The girl gets out of the room in a sudden dash just as man number two gets the door open. The two chase around the outside of the room together, followed by man number one. But they finally elude him and get away. The first man goes back and tries to open his door, but he is so blinded by rage and frustration that he cannot open it. So he butts it open and in a really mad dash around the room he breaks first one room and then the other.
>
> (Heider and Simmel 1944: 246-47)

A few sociological points: the use of "girl" should be read in the context of the time that the experiment took place -1944. Also, the fact that all of the participants were female undergraduates is a product of the time: a large proportion of males aged 19-24 would have been serving in uniform. Likewise, we should not be surprised that all of those interpretations that ascribed characters to the shapes described them as two males pursuing one female. In the United States of the 1940s it would simply not be "the done thing" for two females to pursue the same male aggressively to the point of fighting over him.

For Heider and Simmel's purposes, the interesting words in this description are the *action verb phrases*: "have a fight", "follows", "makes several approaches" and so on. These are taken as evidence of a propensity to describe the film in terms of the intentional actions of the *actors* – even though these "actors" are nothing but simple two-dimensional geometric shapes. To make sense of the actions, the participants assigned anthropomorphic characters

to the shapes – "the girl", "man number one" and "the second man". Incidentally, note the paucity of *mental state* ascriptions: "a man has *planned*", "She apparently does not *want* to be with the first man," and, at a stretch, "blinded by *rage* and *frustratio*n" are the only references to what the "actors" might be thinking or feeling in the whole piece and *they are elaborations rather than being integral to the flow of the action*.

For the purposes of the present chapter the most interesting features of this subject's description are that:

i) The piece is written in the form of a sequence of temporally located events (dictated by the sequence of the film) in which each event follows on as a consequence of the preceding event.
ii) Each consequence is (at least in part) determined by an actor's understandable response to what occurred immediately prior.
iii) The resolution has the "rejected suitor" taking out his impotent rage on inanimate objects – unable to have further impact on the course of events as they involve the other two characters.

What this describes is a *narrative*. In this chapter I will contend that just as we seem unable to avoid describing apparent behaviour in terms of the *intentions* of the participants in that behaviour we cannot help but to situate action *in the form of a narrative*. Again, judging an action as *intentional* need not imply that the actor is in possession of any specific thoughts, propositional attitudes or similar mental states, only that it is not *accidental, incidental or otherwise unintentional*: see Austin (1979b) and *passim* for an elucidation of this distinction. *We are compelled by certain features of our cognitive engagement in the world to order events that involve any kind of intentional agents into narratives*.

4.2 The Ubiquity and Definition of Narratives

Narratives are everywhere. As well as the obvious fictional narratives – novels, epic poetry, fairy tales, films, television and radio drama, comedy and soap opera and so on – we are surrounded every day by narratives in the form of news stories (contemporary culture is probably the most news-saturated there has ever been) through the print and broadcast media. Then there are the myths, legends and religious texts on which culture is built – all are in the form of narratives. Jokes and anecdotes are mini-narratives with slightly different intentions – the first to make us laugh, the second to illustrate a point or to support a

contention. Gossip, excuses, testimony and even documentary use narrative form to different ends. Then we have role-playing games, of the computer or paper variety. It is difficult to imagine how history would even be possible without narrative to convey the sense of a succession of events in a palatable, understandable form. The media through which narratives are transmitted have developed in parallel with our technology. From word of mouth to calligraphy, from block printing to the internet, all have been used for people to tell each other stories. Even pictures can be used to tell a story (Abbott 2008: 7).

We use narratives "… almost from the moment that we begin putting words together. As soon as we follow a subject with a verb, there is a good chance that we are engaged in narrative discourse" (Abbott 2008: 1). Much childhood play comprises making up and acting out stories of one kind or another. For Herman (2007: 17), narrative "can be viewed not just as a means of artistic expression or a resource for communication but also as a fundamental human endowment." Every culture that we know of uses narrative: "narratives are everywhere that humans are" (Abbott 2008: xv). We have written narratives from every literate culture whose script has been deciphered – bar those whose use for "writing" was limited to accounting.[79] Given the contemporary prevalence of narratives in non-literate cultures and the fact that some of our own greatest stories – such as the works of Homer – were originally orally transmitted, it is reasonable to presume that narratives played a significant role in the lives of people for almost as long as we have had language (Nunan and Choi 2010). It has even been suggested that narrative is central to *all* human communication (Fisher 1984).[80]

It would be surprising if something this ubiquitous did not have considerable effects on its users. I will return to those effects and to the psychological purposes to which narratives are put shortly. Firstly, we should try to pin down what "narrative" means.

Ryan (2007: 31-32) laments that "Asking people to decide whether or not a text is a story is one of those artificial situations in which results are produced by the act of investigation," and that "'Narrative' is less a culturally recognised category that influences are choices of

[79] In some senses even a record of the form "so and so produced such and such an amount of grain in a particular year" is a kind of narrative. Especially when combined with a comparative (more or less than the previous year) and a *reason* for the difference (in that year, the rains failed). It is not hard to see how narrative writing developed from record-keeping scripts (Olson 1994; 65-66).

[80] Fisher's "Narrative Paradigm" has important parallels with the thesis advanced in the present chapter, albeit from a different perspective.

reading, viewing, or listening materials than an analytical concept designed by narratologists." It is difficult to settle on a definition of "narrative". Possibly, the very ubiquity of stories offers too many opportunities for counterexamples for any definition that reduces the category to a set of individually necessary and jointly sufficient conditions to be satisfactory. This has not prevented some scholars from attempting a definition.

Abbott (2008: 1) suggests that the fundamental criterion for a piece of language to qualify as a narrative is that it must comprise a *representation of an event*. "Event" can be misleading. The kind of "event" that Abbot describes as being essential to a narrative can include *nothing happening* – provided that nothing happening has significance for the recipient (reader, hearer, viewer etc.) Events, then are temporally distinct occurrences that have some significance or meaning.

In further analysis, Abbott (2008) and Herman (2009b) concur that every narrative comprises two distinct parts: the content that is relayed by the narrative and the mode by which that content is to be transmitted. The content part is synonymous with *story*. The mode of transmission is what narratologists call the *discourse*. These are clearly differentiated when one considers that the same story may be relayed by any number of different discourses – as when one of Shakespeare's plays is adapted into ballet, a narrative form without words (Prokofiev's *Romeo and Juliet*, for example).

This distinction between story and narrative discourse, between what is being told and the telling might be "…arguably, the founding insight of the field of narratology" (Abbott 2007: 36). It is analogous to the linguistic distinction between *signified* and *signifier* (Saussure 1916/2011). However, Abbott cautions that the distinction has been dogged by "two notable controversies":

> One is the question of whether it is a real distinction at all since all we ever know of story is what we get through discourse. Story seems to pre-exist its rendering (note how often stories are narrated in the past tense) yet … the rendering also seems to generate the story, which would make it follow rather than precede the discourse. The other controversy is closely related to the first and involves the repeatability of the story. If the story has a separate existence such that it can be rendered in more than one way and even in more than one medium, how do we know it is the same story when we see it again?

(Abbott 2007: 41)

Both worries are concerned with how a story is altered by – perhaps even dependent on – the telling. To answer the concern about how we can say that the same story is performed by the RSC at Stratford declaiming iambic pentameter as is danced by the Royal Ballet at Covent Garden to Prokofiev's score, Barthes (1988) developed a formal distinction between the *nuclei* and *catalysers* of story, which Herman (2007: 13) characterises as the *core* and *peripheral* elements. Change the core elements and you have a different story altogether (such as a version of *Romeo and Juliet* ending with the young lovers living happily ever after). The story is preserved so long as the mode of telling affects only the peripheral elements. Much work in the analysis of narrative has been dedicated to the reduction of core elements into a set of foundational *archetypes* of story (Frye 1951). Abbott's concerns are over whether altering the discourse is to alter the core features of the story.

This is relevant to the present discussion because when we consider a narrative that describes an action in terms of the mental states of the actor, the question arises as to whether those mental states are *core or peripheral elements of the story*. Could the same story be told, with its core elements preserved, if different mental states were attributed or if the telling relied on other *reasons* for the action (i.e. reasons not connected to the attribution of propositional attitudes at all). I will contend that so long as the reasons for the action (such as situational attributions) were acceptable, we would still have the same story; thus *propositional attitude ascriptions are, at best, peripheral elements in story construction*.

For Abbott (2008: 14), the defining characteristic of a narrative is the "… representation of an event or series of events. … Without an event you may have a 'description', an 'exposition', an 'argument', a 'lyric', some combination of these or something else altogether, but you won't have a narrative. 'My dog has fleas' is a description of my dog but it is not a narrative because nothing happens. 'My dog was bitten by a flea' is a narrative. It tells of an event.

Whether one event is sufficient for a narrative or any narrative needs to describe a sequence of events is, Abbott admits, controversial. As is whether we should require the events covered by the narrative to be causally related. Abbott's claim is only that the representation of a single event is "… the key and it produces the building blocks out of which the more complex forms are built" (Ibid: 14).

In Wittgensteinian mould, (Ryan 2007: 30-31) admits that any definition of "narrative" is likely to be a *fuzzy set* in which we make a judgement that a particular text or utterance is a narrative on the basis of a tacit comparison with a number of *prototypical* cases (cf. Rosch and Mervis 1975). So we have learned, in infancy, what stories are like and when we are told something we compare it to these examples and decide to accept it as a narrative (or not) on the basis of *how closely it resembles (or does not resemble) the prototype examples of narrative that we have stored in long-term memory*. This foreshadows my suggestion of a narrative heuristic, in section 4.8 of this chapter.

As an illustration, consider which of these statements is a satisfactory narrative:

i. The moment she heard about Kepler's laws of motion, Amanda burst into tears.

ii. The moment she heard about her mother's illness, Amanda burst into tears.

I contend that our criterion for evaluating these as narratives rests on our ability to understand, in a broad "folk psychological"[81] sense, Amanda's reaction. Some physics undergraduates might empathise with the first without elaboration, but I believe that we would all agree that the relationship in the second example is more universally *acceptable*. The first would require more information about the actor, Amanda, and her circumstances which *might* follow: my point is only that the second narrative is complete within itself. Each sentence describes two events: Amanda hearing about… and Amanda bursting into tears. And each describes the relation between the actor and each of those events: the first event changes Amanda's state in such a way that she instigates the second event. Unless we know that Amanda has a particular revulsion for 17th century physics, the first does not describe a reaction that we would immediately judge to be reasonable. Knowing our usual assumptions about people's relationship with their parents, the second strikes us as straightforward, understandable and acceptable. Unless it were to read:

> The moment she heard about her mother's illness, Amanda burst into tears *of joy*.

[81] Not necessarily a "belief-desire psychology".

In which case we would find it understandable *only with additional information*. Perhaps the mother's illness proved to be less serious than had been feared. Perhaps Amanda despises her mother. Perhaps Amanda suffers from some affective disorder in which her reactions are the reverse of those we would expect. There is a distinction here between *acceptability* and *truth* that is central to this discussion. All three of these sentences could, conceivably, be true. Our judgement that the relationship in the second example is the most readily understandable does not depend on our judgement that it is the most true, or even the most likely to be true. It suggests only that, in the presence of certain universal background information (child-parent relationships), we find it understandable without elaboration. Similarly, we might find action descriptions featuring "belief" and "desire" in traditional roles (such as the "shopping for milk" example) acceptable as narratives only because these are the default roles that such terms are presumed to fulfil in our culture.

4.3 Narrative Psychology

Narrative theorists, some psychologists and some philosophers have suggested that a facility with narratives plays a central role in how individuals relate to the world. Ricoeur (1984: 3), for example, suggests that it is through narrative that physical time is packaged in to "human time". Abbott (2008: 3) accepts that narratives play a variety of roles in human life and discourse but "…if we had to choose one answer above all others, the likeliest is that *narrative is the principal way that our species organises its understanding of time*" (italics in original).

As part of outline of the way that he wants to define narrative, the narratologist David Herman offers:

> Narrative, in other words, is a basic human strategy for coming to terms with time, process and change - a strategy that contrasts with, but is in no way inferior to, "scientific" modes of explanation that characterise phenomena as instances of general covering laws. Science explains the atmospheric processes that (all other things being equal) account for when precipitation will take the form of snow rather than rain; but it takes a story to convey what it was like to walk along a park trail in fresh-fallen snow as afternoon turned to evening in the late autumn of 2007.
>
> (Herman 2009b: 1)

Narratives allow the individual not just to order events in time or take control of time but also to imbue moments in time with *meaning* at the scale of human experience. Reviewing scientific investigations of the role of stories in action-descriptions Herman (2009a: 56), writes:

> Stories, this research suggests, are a primary technology for making sense of how events unfold in time, one that helps reveal how actions arise, how they are interrelated, and how much salience they should be assigned within a given environment for acting and interacting.

Major alterations to the portrayed temporal relationships can significantly shift the *meaning* of the overall narrative. In this sense it is a *core element* (Herman 2007: 13) or *nucleus* (Barthes 1988) of the story in a way that propositional attitudes need not be.

Bruner (1990, 1991) laments that in its quest for scientific respectability Psychology has lost sight of its true subject matter. He claims that, in thrall to **positivistic, empiricist** ideals, psychologists have tended to ignore the "construction of meaning" (Bruner 1990: 4). The psychologist's central task, in Bruner's view, is to understand how people interpret the worlds that they inhabit and to develop models of how to "interpret their acts of interpretation" (Bruner 1990: xiii). The move to computational analogies has diverted the emphasis of enquiry away from any concern with meaning: *information is "indifferent with respect to meaning" and so, he argues, information should not be the principal concern of a discipline in which meaning is paramount* (Bruner 1990: 4).

It is important to note the distinction that for those philosophers who promote the belief-desire model, a belief is the fundamental unit of *information*. For Bruner, a belief is *one unit of meaning*, in the sense of providing one route through which the individual can relate, understand and participate.

Folk psychology, in the shape of the ascription of beliefs and desires, is described by Bruner as the *lens* through which human action is interpreted. It is "…a culture's account of what makes human beings tick" (Bruner 1990: 13). This is another important distinction. Belief-desire psychology is culturally dependent, according to Bruner. He recognises that different cultures could use different models to achieve similar ends. Culture, alongside meaning, is another theme that Bruner feels has been neglected by Psychology's search for universals. He writes:

> It is man's participation in culture and the realisation of his mental powers through culture that make it impossible to construct a human psychology on the basis of the individual alone.
>
> (Bruner 1990: 12)

Because meaning, the central concern of psychology for Bruner, can serve a purpose only if it is shared, psychology must investigate the mechanisms through which meanings become shared – for example, of how a notion like the belief-desire law becomes embedded within a culture.[82] All meaning is, in Bruner's view, culturally dependent and mediated by shared transactions.

Bruner does not expect a folk psychology and its entities to figure in *causal accounts of action*. "Real causes," he writes "may even not be accessible to ordinary consciousness" (Ibid.: 17). This echoes the views of many contemporary social psychologists – for example Wilson (2002). Bruner argues that:

> Antimentalistic fury about folk psychology simply misses the point. The idea of jettisoning it in the interest of getting rid of mental states in our everyday explanations of human behaviour is tantamount to throwing away the very phenomena that psychology needs to explain. It is in terms of folk psychological categories that we experience ourselves and others. It is through folk psychology that people anticipate and judge one another, draw conclusions about the worthwhileness of their lives, and so on.
>
> (Bruner 1990: 14-15)

Bruner's argument against elimination of beliefs and desires here, directed at the eliminative materialism of Churchland (1981); or Stich (1983) need not directly challenge the *modest elimination* of this thesis. As we have seen, Bruner *rejects the notion of folk psychological entities as describing causes*. In Part One I show that contemporary psychological accounts that *do seek genuine causal accounts of psychological phenomena and behaviour* also eschew talk of these attitudes. Bruner is suggesting a non-explanatory way to do psychology – one in which causes do not figure and so folk psychology – including beliefs and desires - can. The elimination of "belief" and "desire" from *causal accounts of action* would have

[82] Even if that "culture" is made up only of the community of academic philosophers!

no impact on Bruner's project. Equally, Bruner's approach has no direct impact on the eliminative intent of this thesis, save to underline the point that *folk psychology does not belong in the realm of causal accounts*.

Bruner advocates an approach whereby the explanation of action in folk-psychological terms is a starting point for the psychologist's investigations. Folk psychology's role in the construction of meaning (as he sees it) must be restricted to its proper domain. That domain is one in which narratives play a significant part. Bruner's influence has been to give impetus to approaches in psychology that studies how the construction of meaningful accounts of mental life, experience and behaviour depends on the narratives that we build. This is *narrative psychology*.

Narratives are used to construct autobiographical identity, in the description of action and in the ordering of events in time sequences and with causal connections. McAdams (1993) describes human identity as a life story, a "personal myth" on which the unity, purpose and meaning of a life is built. It is an essentially social construct: although we each play a central role in our personal narratives, the story can have no momentum and no cohesion without the contributions of other actors. The myth, according to McAdams, is one that is built on throughout the life of the individual, developing through the addition of experience which gives rise to new compelling narratives.

Polkinghorne (1988: 1) describes narrative as "…the primary form by which human experience is made meaningful". He also suggests that narrative provides "a framework for understanding the past events of one's life and for planning future actions," (ibid. 11). Narrative psychology is concerned with how individuals use stories to construct and convey meaning and investigates the specific stories and story archetypes that individuals and cultures construct. The narratives at work in constructing meaning include "…personal and social histories, myths, fairy tales, novels, and *the everyday stories we use to explain our own and others' actions*" (ibid: 1, emphasis added). This narrative *paradigm* offers a way to describe human life and behaviour that has its own internal logic: it not only builds on the relationship between situations and the individual, between stimuli and responses but also encompasses "emotions, images, time or perspective that have not been treated conceptually so far [within psychology]" (Polkinghorne 1988).

Another proponent of a narrative psychology, László (2008: 9) argues that the discipline "…entails certain assumptions about the relationship between self, identity and social

structures which are distinct from … more 'traditional' perspectives". Those other perspectives – cognitive psychology and experimental social psychology, for example – "…operate on the basis of 'realist' assumptions that are problematic in terms of the study of the self", including "…the assumption that the self exists as an entity that can be discovered and described in much the same way as can any object in the natural or physical world," (Ibid.)

Robinson and Hawpe (1986) suggest that narratives are constructed and used to assist our understanding by means of a **heuristic process** aimed at arriving at an inference or description that "… creates a fit between a situation and the story schema". *Such a narrative is an explanation only to the extent that the event to be understood can be made to fit the narrative*. This means that rather than judging the acceptability of a narrative account of, for example, the antecedents of an action, on the basis of whether or not we have correctly identified the *causes* of that action, *we judge them against our memory of past stories – organised as pre-existing story schemas – and ask ourselves whether this is a likely story*. No attempt is made to uncover the processes by which the events described might bring about the phenomena under examination, which is the essential causal-explanatory concern of cognitive psychology and social psychology as described in Part One of the present thesis. This contrasts with a central commitment of the belief-desire law, according to which the acceptability of an explanation is judged against the presumed causal-functional roles of specific beliefs and desires.

As with Bruner, what differentiates narrative psychology from other approaches is, for Crossley (2000), meaning and meaning-making, interpretation rather than explanation. This difference "highlights the inadequacy of quantitative 'scientific' methods for the study of self and identity." Narrative psychology, she contends, "… advocates the need to focus attention on human existence as it is lived, experienced and interpreted by each human individual." Significantly for the argument to come, concerning the culturally situated nature of folk-psychological narratives, she also argues that such lived experience is "… inextricably tied up with our use and understanding of the linguistic and moral resources made available to us in the cultures that we're brought up in," (Crossley 2000: 10). These

resources are "made available to us" by means of the stories that we begin to be told as soon as we, as infants, are able to understand language.[83]

For the narrative psychologist, this way of constructing meaning from experience is not so much the imposition of structure on a stream if perceptions and ideas: it is essentially how the individual experiences, how raw *information* is shaped into meaningful thought. As Carr (1986: 61) puts it:

> It is not the case that we first live and then afterward, seated around the fire as it were, tell about what we have done. … narration, intertwined as it is with action [creates meaning] in the course of life itself, not merely after the fact.

The suggestion that a central role in psychology should be given to a concept such as narratives, or even meaning, together with some of Bruner's concerns as highlighted above, might seem to be antithetical to the scientific method. László (2008: 4) would deny this, claiming that narrative psychology:

> …assumes that studying narratives as vehicles of complex psychological contents leads to empirically based knowledge about human social adaptation. Individuals in their life stories … compose their significant life episodes. In this composition, which is meaning construction in itself, they express the ways in which they organise their relations to the social world, or construct their identity.

László demands that narrative is something to be taken seriously by science. Our autobiographical histories and identities are not delivered to us fully formed by the events that we experience. Neither are the psychological causes of our actions directly accessible to introspection. Narrative provides the framework for the construction of autobiographical memory and identity an also for the generation of meaningful *reasons* for our judgements, choices and actions. Although the framework of narrative might be fixed, the content of our narratives is contingent on our enculturation, as I will show.

Underpinning the use of narratives in the psychological investigation of the self is this idea:

[83] Perhaps even earlier than this.

> The basic principle of narrative psychology is that individuals understand themselves through the medium of language, through talking and writing, and it is through these processes that individuals are constantly engaged in a process of creating themselves.
>
> (Crossley 2000: 10)

Citing the anthropologist Geertz (1979), Crossley agrees that the view of the self or person that is endemic to the Western (or European) tradition is not typical of the world's cultures. As Geertz describes it, the habit of seeing the person as:

> … a more or less integrated motivational universe, a dynamic centre of awareness, emotion, judgement and action, organised into a distinctive whole and set contrastively against other wholes and the social and natural background is … a rather peculiar idea within the context of the world's cultures.
>
> (Geertz 1979: 229)

This leaves something of a "chicken and egg" question: does the Western conception of identity emerge from the prevalent kinds of narrative in western culture or do our narrative traditions proceed from a conception of the person as "a dynamic centre"? For our purposes here this debate can be left aside. It is significant that our prevalent narratives either shape or are shaped by our understanding of the individual as an "integrated motivational universe".

Perhaps the most influential *philosophical* examination of narrative identity is that of Ricoeur (1984, 1991). He contends that the foundations of two principal components of identity – unity and permanence – rest on the narratives through which the self is understood. It is through such narratives that meaning and temporal structure is given to the stream of experience. It is only through narrative's inherent temporal structure (beginning-middle-end) that humans are able to appreciate time at all. The social-constructivist view of identity expounded by Ricoeur and others might be linked to narrative psychology and summed up thus:

> A narrative conception of identity implies that subjectivity is neither a philosophical illusion nor an impermeable substance. Rather, a narrative identity provides a subjective sense of self-continuity as it symbolically

> integrates the events of lived experience in the plot of the story a person tells about his or her life.
>
> Ezzy (1998: 239)

Narrative psychologists have sought to bring this concept under experimental examination. In their review of the empirical literature, McAdams and McLean (2013) describe a series of experimental investigations concerning the impact of specific narrative constructions of identity on development and adaptation to changing life circumstances. They draw attention, in their conclusion, to the effect that differences of culture, different "menus of images, themes, and plots for the construction of narrative identity" can have on how individuals, embedded within a culture, come to build the narrative identity that they do. The remaining challenge, they suggest, is to extend these investigations to comparisons between cultures and between narrative styles.

This inevitably leaves open the idea that if a person can make changes to their habitual narratives they can make radical changes to their self-conception. This is precisely how narratives are used in therapeutic approaches.

4.4 Therapeutic Narrative Psychology

The underlying principle of applying narratives to therapeutic approaches is that *the stories we tell ourselves directly influence our picture of the world*. If the picture generates problems, the therapist can help the subject to challenge it. Changing the narrative (or stepping outside of it) generates a more empowering world-picture:

> The therapist helps clients articulate and bring to language and awareness the narratives they have developed that give meaning to their lives. The clients are then able to examine and reflect on the themes they are using to organise their lives and to interpret their own actions and the actions of others. The reflective awareness of one's personal narrative provides the realisation that past events are not meaningful in themselves, but are given significance by the configuration of one's narrative. This realisation can release people from the control of past interpretations they've attached to events and open up the possibility of renewal and freedom for change.
>
> (Polkinghorne 1988: 188)

Patients are firstly encouraged to uncover and to examine the narratives that they use to describe their present situation, their past experiences and their future expectations. The therapist guides them to notice the effects that their narratives are having on them. For example, a subject who constantly rehearses their mistakes might have settled on a narrative that reinforces the view "everything I do goes wrong" and so be reluctant to try anything new or to welcome change. Often, as Pennebaker (2004) notes, *simply acknowledging the hold that negatively charged narratives have had on them leads people to positive effects*.

I make no apology for the obvious parallels between this and the diagnostic-therapeutic intent of the present thesis.

Many *talking therapies* have narrative elements – including *cognitive behavioural therapy* or *CBT* (A. T. Beck 1979), one of the most widely used interventions funded by the National Health Service in the UK (Stiles et al. 2008). For an example of a narrative therapy at work, however, I will describe the *story editing* approach pioneered by Pennebaker (1997).

Wilson (2011), in reviewing this method, compares story editing favourably with more traditional approaches to recovery after traumatic experiences. One traditional approach involves recounting the facts of the traumatic event and, under the guidance of a therapist, reliving experienced emotions. This *critical incident stress debriefing* technique is widely used. However, Wilson (2011) claims that some evidence suggests that this is at best ineffective and might even be harmful.

Story editing, in contrast, involves the subject developing a narrative account of the traumatic event, reshaping the emotional content to their present needs. Wilson argues that:

> Our interpretations are rooted in the narratives we construct about ourselves and the social world, and sometimes … we interpret things in unhealthy ways that have negative consequences. We could solve a lot of problems if we could get people to redirect their interpretations in healthier directions.
>
> (Wilson 2011: 9)

Story editing is defined as "A set of techniques designed to redirect people's narratives about themselves and the social world in a way that leads to lasting changes in behaviour," (Wilson 2011: 11). Subjects are invited to write about their traumatic experiences repeatedly for a fixed time each day for the duration of their therapy. Each time they do so, they are

given licence to diminish some elements and enhance others. Pennebaker (1997: 95) suggests that this ensures that:

> Over time, individuals who are writing about an event become more and more detached. They are able to stand back and consider the complex causes of the event and their own mixed emotions. Perhaps by addressing the trauma multiple times, people's emotional responses become less extreme. In other words, repeatedly confronting an upsetting experience allows for a less emotionally laden assessment of its meaning and impact.

For the purpose of testing the technique, to find out whether it really could accelerate "adjustment to ongoing life transitions" (Ibid.: 79-80), Pennebaker sought experimental subjects dealing with a major life-upheaval. As a university academic, the answer was on his doorstep; college freshers, many of whom would be away from home for the first time in their lives, often in an unfamiliar locale and always in a highly pressured environment.

Pennebaker (1997: 80) describes his procedure as follows: "In order to learn if we could accelerate coping, we asked about 130 entering freshmen to participate in an experiment that dealt with 'writing and the college experience'." These students wrote for 20 minutes each day for three consecutive days. The subjects were allocated to one of two experimental conditions by the flip of a coin. Those in the first condition were invited to write about some superficial topics, while students in the second condition were told:

> For each of the writing sessions, I want you to let go and write about your very deepest thoughts and feelings about coming to college … In your writing, you might want to write about your emotions and thoughts about leaving your parents, about issues of adjusting to the various aspects of college … or even about your feelings of who you are or what you want to become.

One surprising result was that a considerable number of students in the second condition took the opportunity to record their thoughts about genuine traumas that they had suffered. Some writings included "… suicide attempts, family violence, rape … basically the same things that I had read when people had been asked to write about traumas" (Pennebaker 1997: 81).

In all cases, the subjects who wrote about their college experience had fewer visits to the doctor and reported illnesses than did those in the control condition. There was no

discernible difference between those who took part in September and those who joined the programme in December.

Starting university, for all the upheaval, is a relatively benign circumstance. Pennebaker wanted to try out these techniques with a group whose experience must be among the most extreme examples of trauma in living memory.[84] Through the Dallas Memorial Center for Holocaust Studies (DMHC), Pennebaker made contact with sixty survivors of the Nazi death camps. More than seventy per-cent of these people reported that they had never spoken of their experience with anyone – including close family. The motivating question behind this research was this: would it be possible to accelerate the coping process even for this unique group?

The subjects were invited to speak of their experience in the safe environment of a psychological interview. While they spoke, their skin conductance and heart rates were monitored. This would establish a base rate for the emotional intensity of their telling the tale. One year after their initial interview, each subject was invited to a health assessment.

> Using the skin conductance data, as well as ratings of the content of each survivor's testimony, [we] could define each survivor as a high discloser, a midlevel discloser or a low discloser. High disclosers were people who, when they told of the personal traumas that they had suffered, remained physiologically relaxed. Low disclosers, on the other hand, exhibited biological signs of increased inhibition and tension when disclosing traumatic events. Overall, we found that high and midlevel disclosers were significantly healthier in the year after the interview than before it.
>
> (Pennebaker 1997: 86)

Pennebaker acknowledges that there are individual differences in coping strategies: some people my do better by "bottling up" their emotions than would others. Notwithstanding this, he is convinced that in the case of people for whom trauma is having a lasting impact, writing, rewriting and *opening up* about the experience is beneficial. "Other studies in addition to the holocaust project point to the same conclusion. If you are currently living with a trauma from years gone by, writing or telling about it can help you get past it. … If

[84] There were many more people for whom these events were "living memory" in the mid 1990s.

something horrible happened to you five years ago and you are still living with it, writing about it will likely help," (Ibid.: 88)

Pennebaker suggests that "People have a basic need for completing and resolving tasks," (Ibid: 90) and that "We are often so intent on finding meaning in an event that we become irrational," (Ibid.: 92). As evidence of this latter tendency, Pennebaker cites the prevalence of *victim blaming* and the *just world hypothesis* (Gilovich et al. 2006: 360). Following a traumatic experience "we naturally search for meaning and completion to events that we know at some level don't have meaning and can never be resolved" Pennebaker (1997: 92)

By writing down their experience, or telling the tale of what happened – preferably repeatedly, Pennebaker maintains that they construct meaning from the jumble of experience. *Because narratives must conform to the beginning-middle-end pattern they impose a resolution, of sorts, on the previously unresolvable*. Perspectives are changed and, consequently, emotional bearing managed. Pennebaker writes "… repeatedly confronting an upsetting experience allows for a less emotionally laden assessment of its meaning and impact" (Pennebaker 1997: 95). Note that this does not imply any altered understanding of the *causes* of a traumatic event. By making the story their own, trauma survivors are better able to deal with what happened. This is supported by the contention that **acute stress** or **post-traumatic stress disorders** can be understood either as a break with an established narrative, a gap between reality and the narrative that we use to explain and understand it or as the absence of a suitable narrative to accommodate an experience (Brewin 2001; Davey 2008: 490; Palgi and Ben-Ezra 2010).

A significant correlation between writing style and health benefits was that *overall narrative coherence seemed to be a positive indicator*:

> We realised that the people who benefitted from writing were constructing stories. On the first day of writing they would often tell about a traumatic episode that simply described an experience, often out of sequence and disorganised. But day by day, as they continued to write, the episode would take on shape as a coherent story with a clear beginning, middle and end.
>
> (Pennebaker 1997: 104)

The imposition of narrative structure showed the highest degree of correlation with improved health. By ordering the raw and anarchic elements of a traumatic experience –

what might be characterised as a violation of the *expected narrative* (Crossley 2000) – the individual is more able to come to terms with it.

> We need to construct coherent and meaningful stories for ourselves. Good narratives or stories, then, organise seemingly infinite facets of overwhelming events. Once organised, the events are often smaller and easier to deal with.
>
> (Pennebaker 1997: 103)

The foregoing section underlines the importance and the power of narratives in constructing our world-view. Could not the philosopher, however, object on the grounds of *truth*?

4.5 Narrative Truth versus Historical Truth

> But I did not get my picture of the world by satisfying myself of its correctness; nor do I have it because I am satisfied of its correctness. No: it is the inherited background against which I distinguish between true and false.
>
> Wittgenstein, *On Certainty*, §94

If the standard is one of **correspondence** with historical facts, the *truth* of the constructed account – at least in so far as words are chosen to minimise negative emotional consequences – might be doubted. How far should the therapeutic narrative be allowed to diverge from facts about the patient's history? As Crossley (2000: 61, emphasis added) asks:

> If certain psychotherapeutic techniques encourage us to imagine alternative possibilities and imaginatively rewrite our stories, to what extent are they committed to historical truth? Does *narrative truth*, construction of a pleasing, coherent and persuasive story, take precedence?

The criterion for *narrative truth* is not that the story *corresponds to the historical events*. It is perhaps more important that the narrative serves the purpose of constructing a *useful meaning* for the individual. *Usefulness* is dependent upon circumstances or the occasion for which the narrative is deployed. The notion of narrative truth and the construction of meaning are thus compatible, just as each is incompatible with a correspondence standard of truth.

Bruner (1990: 61) maintains that "We interpret stories by their verisimilitude, their 'truth likeness', or more accurately their 'lifelikeness'." This distinction, between "verisimilitude" or "lifelikeness" and causal-explanatory (or historical) truth is key to an appreciation of how we can judge a narrative with a particular content **acceptable** without committing to the historical truth of the implied relations between the entities that it contains. *A folk psychological reason for an action can ring true without being available as a causal explanation of that same action*. Bruner (1990: 118) again:

> … there are no causes to be grasped with certainty where the act of creating meaning is concerned. Only acts, expressions and contexts to be interpreted.

This implies a different interpretation of truth. Spence (1982) applied just such a distinction to **psychoanalysis**. His project was to differentiate the truth of descriptions of events from the unique perspective of a patient and the truth of events – potentially the same event – from the objective, historically accurate standpoint. He describes this as the difference between **narrative truth** and **historical truth**. Events might have meaning and resonance for the individual patient; they might even be revealing of the objective ways in which that patient behaves. However, the narratives that the patient constructs to describe their role in the events and the reasons for their behaviour do not necessarily objectively describe actual events related in the precise chronological order that they occurred. By putting things into words, constructing our own meaningful narratives from the raw material of episodic memory *we simultaneously construct our own truths*.

> The evidence from the past that emerges in the course of our clinical work may be used to confirm our search [for formative events], but it has fallen under the shadow of our construction, whatever it happens to be. … The construction not only shapes the past – it becomes the past in many cases because many critical early experiences are preverbal and therefore have no proper designation until we put them into words.
>
> (Spence 1982: 175)

Spence points out that *narrative fit* (cf. Bruner's *verisimilitude*) is often all that the psychoanalyst[85] in the clinical setting *or the lay person in the everyday setting* demands in order to accept a particular narrative as true.

> A particular clinical event … may seem to clarify the unfolding account of the patient's life history so precisely that both patient and analyst come to the conclusion that it *must* be true. …narrative fit is usually taken to be conclusive, and if a piece of the past completes the unfinished clinical picture in just the right way … then it acquires its own truth value and no further checking is necessary.
>
> (Ibid.: 181)

In common with Bruner's psychology, Spence's conception of psychoanalysis is of a discipline concerned with the interpretation of meanings. These are uncovered by *examining the constructed narratives that the subject uses to make sense of their experience rather than seeking evidence for the historical causes of present psychic disturbances*.

One of the pioneers of narrative psychology, Sarbin (1986) also cautions against regarding the *interpreted, socially constructed truths* of narratives – including, explicitly, folk-psychological narratives – as true causal-explanatory accounts of action. He suggests that we can regard narratives of action as guiding patterns for life and conduct that become useful only when used as a pre-existing scaffolding for the interpretation of events. Sarbin suggests that we ask ourselves "what kind of narrative is this like?" and impose that structure on the received event. When the events are new, unexpected and resist explanation according to the situational or personal schemas of ordinary attribution (see Chapter 3) we construct a narrative that fits by modifying one from our stock of ready-made structures. This is, I argue below, indicative of the application of some kind of *narrative assessment heuristic*.

According to Sarbin, narrative is the root metaphor of human psychology, the fundamental scaffolding for our experience of the world. He writes "… human beings think, perceive, imagine, and make moral choices according to narrative structures" (Sarbin 1986: 8). Narrative is "a fruitful metaphor for examining and interpreting human action," (Ibid.: 19).

[85] Spence's concern is with psychoanalysis but for my purposes it is safe to assume that the same consideration applies to other psychotherapeutic practices employing narratives.

It would thus be a *mistake to regard a narrative, constructed to provide a ready-made scaffolding for an action or an experience of action for a causal-explanatory account of that action*.

In the context of psychotherapeutic efficacy, Spence is adamant that, although recognising the difference, we should allow the narrative to take the lead. What is important, he argues, is the *meaning* that the patient has constructed through narrative, rather than historical accuracy. The therapist's role is to make changes to the narrative that bring about helpful changes to the meaning, diminishing negative associations and augmenting positive interpretations. In this way, the patient might be liberated, if not from the history (which we assume to be immutable) but from the *meaning* that they have attached by means of their narrative. The narrative psychologist requires only that our narratives of those episodes, together with the overriding narrative through which episodes are interrelated and given structure, are sufficiently *coherent* to offer a *narrative truth*. The contention of therapeutic narrative psychology is that certain mental pathologies result either when those attached meanings have negative connotations or when coherence breaks down.

In answer to the question with which this section began, scientists, such as the psychologists featured in Part One of this thesis, should be concerned with *historical truth*, in which causes can be identified and categorised. Since belief-desire psychology is best suited to the realm of *narrative meaning construction*, it is not suitable for this purpose. Philosophers should be aware of the distinction and ensure that belief-desire accounts are restricted to their proper, non-causal, domain.

This finds philosophical resonance, I suggest, with Austin (1962) when he urges philosophers to pay attention to the roles that utterances play in discourse:

> It is essential to realise that 'true' and 'false', like 'free' and 'unfree', do not stand for anything simple at all; but only for a general dimension of being a right and proper thing to say as opposed to a wrong thing, in these circumstances, to this audience, for these purposes and with these intentions.
>
> (Austin 1962: 145)

4.6 Narratives With or Without Mental State Terms?

Surprisingly for those who regard belief-desire attributions as a fundamental way that humans organise stories, explicit ascriptions of mental states tend not to feature in narrative storytelling, particularly literary narratives. One, perhaps the main, narrative archetype is the *quest,* in which the agent driving the story (the protagonist) seeks to restore equilibrium to a world that has been disrupted by some force beyond the protagonists control. This force is known by film story-writing teacher Robert McKee as the *inciting incident* of story (McKee 1999: 181).

This is coterminous with the "disruption or disequilibrium" that Herman (2009b) suggests is introduced into the character's story world and from which they seek to recover their balance – even if the resolution of the story is an entirely new state of affairs.

It is *possible* to cast this quest model of narrative in terms of belief-desire psychology. We could insist that, after the inciting incident, every action the protagonist takes is one that they *believe* will bring about their *desire to restore balance to their world*.

However, making any propositional attitude states of the protagonist explicit is not helpful to the storyteller. Recipients of narratives find that the inciting incident or upset that the protagonist suffers is a sufficiently powerful situational stressor (cf., references to **situational attribution** in Chapter 3) to lead to whatever action the protagonist takes. Their choices of action will reveal their underlying character or personality traits (see **personal attribution**, again in Chapter 3). Revealing character through action choice in this way is the origin of the oft-cited suggestion that would-be creative writers concentrate on *showing rather than telling their story*. Author Chuck Palahniuk recommends eschewing "thought verbs" (believe, understand, desire, hate etcetera) altogether (Palahniuk 2013). According to McKee (1999: 144-45), narratives that explicitly reveal a protagonist's intentional mental states are unsatisfying to the reader, viewer, listener or hearer: this way of making the inner life of characters explicit and serves only to remind the audience that what they are experiencing is an artificial construct.

As Hutto (2008b), Ratcliffe (2006, 2007) and Reddy and Morris (2009) point out, in real life, most of the time we have no need to posit a relationship between propositional attitudes and behaviour in order to make sense of what a particular agent does under which particular circumstances. This also seems to be the lesson of attribution theories in social psychology

(Chapter 3, this thesis). All that is required is the story of *who* did *what* and *when.* Combining the specific character traits of the agent with the prevailing circumstances and the temporal position relative to other relevant events builds a *coherent, meaningful and acceptable narrative*.

The fact that propositional attitudes are usually omitted does not refute the contention that the belief-desire law underlies action choice. However, it does suggest that we are fully equipped to make sense of actions without having the propositional attitudes that the agent possessed at the time spelled out or imposed on the protagonist.

The author who is committed to FP-style explanations of behaviour might, conceivably, advert to explicit belief-desire language in making the "inner life" of his characters obvious: this is a characteristic of much bad writing. The objective here is to show that we can understand action without such exposition.

Evidence from our oldest literary traditions and from at least one other culture suggest that *an appreciation of people as minded, intentional actors is not dependent on a mental-state vocabulary.*

An objector to my stance thus far in this chapter might suggest that because narratives of action featuring "belief" and "desire" ascriptions have appeared "throughout our literary history" they must have some foundation in truth. One way that this objection was put directly to me is that "Belief and desire action explanations are even in Homer!"

It would be possible to argue against this point on the grounds that *just because we can cite instances of the description of action in terms of causal beliefs and desires does not guarantee their truth*. It could be that all such narratives are never more than narratively true, that they might provide a certain meaning (perhaps that the agent is the author of their own actions) but this does not establish the causal relationship between beliefs and desires and actions. This would be one way of approaching the objection. It would be unlikely to convince the interlocutor who maintains that the ubiquity and *antiquity* of such action descriptions must have originated in some recognition of their underlying *truth*.

More fatal to the objection, however, is that the assertion "they are even in Homer" is, objectively, *not true*.

> The most striking feature of the Homeric conception of mind is that *they had none*. There is no evidence of a concept of mind as distinguished from body and there is an absence of such terms as 'decided', 'thought', 'believed' or 'equivocated'.
>
> (Olson 1994: 238)

David Olson, a professor of applied cognitive science, used textual and historical analysis to investigate the way that advent of writing and the coming of widespread literacy to populations has had a profound effect on our relationship with information and our mental lives. Since the Homeric epics were, originally, orally-transmitted narratives, designed to be memorised and performed to groups of listeners, this would, he theorised, make them significantly different in character to the kinds of narrative that are present in our modern, literate culture.

Citing earlier analytic work (Jaynes 1976) on Homer and on the earlier parts of The Bible,[86] Olson maintains that the kinds of occurrences that might today be described in mental state terms are, instead, *described as bodily sensations*. For example, "feelings or emotions are referred to as the palpitating heart or panting breath or uttering cries," (Olson 1994: 239). If a character in the Homeric epics makes a choice or decision, he is reported as *hearing voices, telling him what he should do*. In the Iliad, Book 19, 100-104 (Homer 1991) we have Agamemnon simultaneously accepting responsibility and passing blame for his having taken Briseis from Achilles:

> But I am not to blame!
> Zeus and Fate and the furies stalking through the night
> They are the ones who drove the savage madness in my heart
> That day in assembly when I seized Achilles' prize
> On my own authority, true, but what could I do?

Olson notes that Bruner (1990, 1991) has postulated that the sense of self emerges from "story-telling" involving the "narrative 'I''. The bards who composed the Homeric epics, before they were committed to writing and attributed to a single semi-mythical author, *seemed to lack this sense of an autobiographical narrative*. "Action, for them, appears to

[86] Given that the first written versions of the Homeric epics are thought to date from the 8th century BCE (Nicolson, 2014) this would make the older parts of the Bible roughly contemporaneous (Jaynes, 1976).

reside in the collective which included the gods, rather than in the individuals and their minds," (Olson 1994: 240).

In the Homeric period, even those ancient Greek terms that are today inevitably associated with mind are used in ways indistinguishable from their bodily effects. The term *"psyche"* (ψυχή), for example, which is today taken as a cognate for "mind" seems to mean nothing more than "life": when people are killed, their *psyche* "bleeds out" of them.

Even today, there are cultures whose language has no mental-state terms such as "belief" and "desire". The Junin-Quechua people of south America speak one of the successor languages to those of the Inca civilisation. This is predominantly an oral tongue, having no indigenous script of its own; although it has been transcribed into the Roman alphabet since the Spanish colonisation of South America. According to an investigation by Vinden (1996) this language has no cognates for "belief" or "desire" nor any words designating mental states. The culture has acquired some mentalistic terms since colonial times in the form of Spanish loan-words. However, these are altogether absent from some of their older (pre-Colombian) folk tales.

This lack does not prevent the Junin-Quechua having a rich narrative tradition. An illustration will show how a non-mentalistic culture tells stories and how this is possible without such constructs.

Here is how Vinden briefly relates one of the Junin-Quechua folk tales:

> The story of the Fox and the Cheese provides an example of such a folk- tale. Briefly, the tale is about a very hungry fox who sees a large, round cheese in the middle of the lake. As the narrator tells us, however, it is really the moon's reflection, which looks like a cheese. When he asks an owl to help him, the owl refuses, saying that it's not a cheese. The fox attempts to swim out and eat it, and drowns. In the Junin Quechua version of the story, there is clear use of the language of appearances to describe the fact that the moon *looked like* a cheese. However, the thoughts and beliefs of the fox and owl are never discussed.
>
> (Vinden 1996: 1708)

Vinden (1996) used this example to describe how differently people from our culture, with their rich panoply of mental-state terms, would retell this story. A literal translation of the

story was read out to a small group (seven) of Canadian graduate students. They were then asked to write the story down as they remembered it. All of these subjects added at least one mental state term to their version of the story.

Vinden also wanted to investigate whether the absence of mental-state terms from the language would hamper the ability of Junin Quechua children to understand the mentality of other people. Groups of Junin Quechua children were set a series of false-belief tasks.

The classic *false belief task* is the "Sally-Anne" test (Baron-Cohen et al. 1985). Subjects (usually children) are shown a scenario acted out with two dolls. One of the dolls, Sally, is in possession of a marble. The experimenter shows Sally placing her marble in a hiding place and then leaving the room. The second doll, Anne, then takes the marble and places it in a different hiding place. The subjects then are asked where Sally, on returning to the room, will look for her marble?

Most subjects over the age of about four years suggest that Anne will expect her marble to be where she left it, in other words she will have a *false belief*. Children under four and autistic adults typically "fail" this task and predict that Sally will go straight to the new hiding place, where Anne has secreted her marble (ibid.). This is taken as evidence of a failure to ascribe false beliefs to Sally. Non-verbal versions of the false-belief task have suggested that non-human primates also lack this ability (Call and Tomasello 1999). Interestingly, pre-verbal infants, using the same gaze-tracking methods, seem to be more successful at anticipating that Sally will look for her marble in her original hiding place (S. C. Johnson 2000).

Vinden's results with the Junin-Quechua children were similar to those expected from children of equivalent age whose native European languages are replete with mental-state vocabulary. This should not be surprising: the story of the fox and the cheese requires some conception of being mistaken, of acting on false information. The finding is in accordance with Olson's comment regarding the Homeric epics that:

> Homeric folk psychology was therefore quite different from ours in at least one way. Like us, they understood lies and deception – the competence that distinguishes children under four years, autistic adults and non-human primates from normal humans – yet they lacked a vocabulary and its corresponding concepts for thinking about the mind.

(Olson 1994: 240)

This suggests that an understanding of deception and mistake and the realisation that people can be subject to these does not rest on the deployment of a term for "belief" or a cognate of "false-belief". The concepts of deception, error and so on are thus potentially independent from a causal conception of "belief". In her conclusion, Vinden suggests that:

> Perhaps, however, we should also cast our nets more widely in a variety of cultures, seeking not so much to discover common achievement on such things as theory of mind tasks, but also looking for ways to explore more generally the extent to which people think of individuals as holding private, interpretable mental states.

(Vinden 1996: 1715-16)

Chapters 2 and 3 of this thesis have shown that that many cognitive and social psychologists do not employ mental-state terms to describe or explain action choice or interpersonal understanding, respectively. Taking the observations of Olson alongside those of Vinden also suggests that *employing mental state terms in narratives of action is culturally dependent and describing action in belief-desire terms is an artefact of our contemporary culture*.

4.7 Narrative Folk Psychology

As suggested in the introduction, two descriptions of belief-desire psychology have vied for the dominant position and spawned a number of intermediate positions. **Theory-theory** has it that people are equipped with an innate theory of the relationship between beliefs and desires and action. **Simulation theory** suggests that we run *off-line simulations* of how we would act if attending to particular belief-desire sets. In recent years these two have been joined by a third distinctive entrant to the debate, the **Narrative Practice Hypothesis (NPH)**, proposed by Hutto (2008b, 2008a, 2009).

Hutto argues that we should adopt some lessons from Bruner (1990) and from narrative psychology if we want to understand how folk-psychological accounts develop and come to seem so plausible as *reasons* for actions.

> Folk psychology just is the practice of making sense of intentional action by means of a special kind of narrative, those that are about or feature a person's reasons.
>
> (Hutto 2008b: 7)

Hutto offers an account of the development of folk psychology as a skill, suggesting that a *mature individual's ability to deploy reasons is allied with and dependent upon that individual's acquisition of competence with narratives*. As primary evidence for this, he marshals research from developmental psychology that suggests that our ability to understand and deploy belief/desire reasons and our ability to construct coherent narratives reach something approaching their mature facility around the same age in most individuals (typically four years of age). His critique of more traditional, theory-based, accounts of how folk psychology works is more conceptual. Defining "folk psychology" as the human facility to make sense of actions in terms of reasons, he claims that the NPH, postulates *"no dedicated inherited inner mechanisms"* (Hutto 2008b: 246) because the practice of attributing propositional attitudes in order to provide reasons for action would be an integral part of the practice of constructing narratives, which he, along with the narrative psychologists and philosophers such as Ricoeur (1984, 1991) and MacIntyre (1981) (see below), take to be a significant developmental stage in its own right. The NPH suggests that narrative competence does everything that we might require a dedicated mindreading mechanism to do.

Hutto is at pains to point out that many of our social transactions do not depend on the attribution of folk-psychological states to others at all. Sometimes, he recognizes, the situation, or some character trait or stable disposition that we attribute to ourselves or others, is sufficient to offer a reason for specific instances of behaviour. He writes:

> It is therefore false to say that without folk psychology we would be bereft of any reliable means of interacting with others. Nor do we call on it that often. Many of our routine encounters with others take place in situations where the social roles and rules are well established, so much so that unless we behave in a deviant manner we typically have no need to understand one another by means of the belief/desire schema. More often than not we neither predict nor seek to explain the actions of others in terms of their unique beliefs and desires at all.

(Hutto 2008b: 3-4)

Even if one accepts Hutto's hypothesis that folk psychological competence develops through the acquisition of skills at understanding, constructing and deploying narratives, could it not be that our mature understanding of the roles of belief and desire in providing reasons for an agent having acted in a particular way is "abstracted away" from narratives? That is to say, we might learn how beliefs and desires work by seeing their roles within certain kinds of narrative but once we have acquired this understanding we are able to deploy them to designate causally efficacious mental states – without the "narrative scaffolding" that allowed the individual to gain an understanding of their roles. Hutto, however, maintains that *the ability to account for actions in terms of reasons remains, throughout our lives, vested in our ability to deploy narratives that feature propositional attitude attributions*.

Hutto's objective is that folk psychology should be immunised from certain philosophical concerns that have dogged the theory-theory and simulation-theory accounts. Of particular significance to the stance taken in the present thesis, he suggests that regarding a belief/desire claim as narratively grounded avoids issues over the *degree to which such claims can be satisfactory as causal explanations* – at least compared with the demands of scientific causal explanations and the resulting doubts over issues of mental causation.

> … at least in some cases the mere citing of the appropriate belief/desire, even when it fits appropriately with the other relevant beliefs and desires and even assuming that it is causally responsible for the action in question, does not suffice to explain an action in the strong sense of making it intelligible. A larger narrative that further contextualises the reason, either in terms of different cultural norms or the peculiarities of a person's history or values, is required for that (that is, if anything can achieve this end).

(Hutto 2008b: 8)

We determine whether a belief/desire claim *satisfactorily explains* an action, this implies, by the acceptability of the narrative. This contrasts with the question of causal determination, which scientist would demand of a mechanistic explanation such as those arrived at by functional decomposition in psychology (Chapter 1).

Ratcliffe (2008, 2009) agrees with Hutto that reasons couched in terms of beliefs and desires are not the only nor even the most usual way in which we understand our own or others' actions. However, for Ratcliffe, any description of folk psychology as the way that ordinary people ("the folk"), free of the constraints or requirements of either scientific explanation or of philosophically rigorous precision, account for actions is either an unwarranted oversimplification of or an unhelpful abstraction ("an uninformative caricature", he alleges) from a rich variety of ordinary practices which may vary by the moment, let alone by the occasion. In his view, describing "our so-called folk psychology requires difficult philosophical work" that, he claims, has not been done (Ratcliffe 2009: 379). He alleges that, for all its strengths, Hutto's NPH suffers from being another attempt by a philosopher to identify and describe the abilities that allow us to deploy belief-desire reasons rather than addressing the pressing and more fundamental question of "…whether 'using belief-desire psychology to predict and explain behaviour' is something that we do." (ibid: 380). This, of course, is part of the rationale for the present thesis.

Also, even if Hutto is right about the fact that competence with belief-desire reasons, on the limited occasions that they are deployed, emerges from our competence with narratives this is entirely consistent with the idea that the content of the belief-desire hypothesis is culturally determined. As such, it would be a feature of the narratives that *we* use, in this culture, today. It need not be an essential part of action choice or interpersonal understanding for all humans, everywhere and forever.

Another influential view of the central role of narratives in the development of our relationship with the world comes from MacIntyre (1981: 54)

> It is through hearing stories … that children learn, or mis-learn, both what a child and what a parent is, what the cast of characters may be in the drama into which they have been born and what the ways of the world are. Deprive children of stories and you leave them unscripted, anxious stutterers in their actions as in their words. Hence there is no way to give us an understanding of any society, including our own, except through the stock of stories which constitute its initial dramatic resources. Mythology, in its original sense, is at the heart of things.

I suspect that Hutto, MacIntyre and even Ricoeur would agree that individuals come to a mature understanding of the world, themselves and each other only once they acquire a facility with narratives. For philosophical accounts of the role of narratives, and for the

narrative psychologist. stories are a tool by means of which individuals impose temporal order on and make sense of the chaotic stream of experience: as rationalisations of what agents do, their value lies in the way that they permit the person who constructs them to impose order on their appreciation of the world of the world – *not to explain it in terms of causes and effects*. The commitment to truth in the everyday narrative of action is not the same as the commitment to causal-explanatory adequacy required by science.

> Using such narrative procedures does not presuppose or entail having a theory of one's own or another's mind. Rather it involves building a model of how *actions* are situated in time and (social) space, and of how they emerge from and impinge upon the larger pattern of actions that constitutes all or part of a person's life-course.
>
> (Herman 2009a: 69)

I have suggested that philosophical folk psychology, belief-desire psychology and its central tenet the belief-desire law constrain philosophical thinking by presenting diverse phenomena as a single *philosophical picture*: a fixed viewpoint that excluded alternatives. Perhaps, in the light of the present chapter, I might employ an alternative metaphor: perhaps we should see philosophical folk psychology as a *persistent* ***narrative*** *of action-explanation, interpersonal understanding of reason giving*. Like a culture's fundamental mythology, it is difficult for anyone in thrall to the world view to admit that there are other ways to relate the phenomena. However, other *narratives of action* are available – both within other cultures and within the specific fields of social and cognitive psychology (chapters 3 and 2, respectively).

> The propositions describing this world picture might be part of a kind of mythology. And their role is like that of rules of a game; and the game can be learned purely practically, without learning any explicit rules.
>
> Wittgenstein, *On Certainty*, §95

4.8 Is There a "Narrative Heuristic"?

Mythology, fairy tales, gossip and news reports: we are drenched in stories from our earliest years. We learn that much of the information that we use to understand the world comes in the shape of a narrative. In this section I want to speculate that the ubiquity of narrative has

an even more insidious effect: whether our choice of whether and what to accept might be influenced by the way the information is presented. I want to ask whether it is possible that we base some of our judgements on a **narrative heuristic**.

One persistent systematic divergence of reasoning from the entailing norm is the **conjunction fallacy**. This is taken as evidence of the use of a one-part heuristic – **representativeness –** for which the decision rule is: *the option with the highest probability [of class membership] is that which is similar in essential characteristics to its parent population and which reflects the salient features of the process by which it is generated*. A classic demonstration of this effect is known as the **Linda case**. Subjects were tested by being offered the following vignette:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> (Tversky and Kahneman 1983: 297)

And were then asked which of the following statements about Linda is most probable:

> Linda is a teacher in elementary school.
> Linda works in a bookstore and takes Yoga classes.
> Linda is active in the feminist movement. (F)
> Linda is a psychiatric social worker.
> Linda is a member of the League of Women Voters.
> Linda is a bank teller. (T)
> Linda is an insurance salesperson.
> Linda is a bank teller and is active in the feminist movement. (T&F)
>
> (Ibid.)

For the three questions of most interest to the researchers, T, F and T&F, 85% of respondents ranked these in this order of probability: F>T&F>T. In a subsequent test, an even larger group of subjects was offered to choose the most probable between just these two possibilities:

> Linda is a bank teller. (T)

> Linda is a bank teller and is active in the feminist movement. (T&F)
>
> (Ibid.: 299)

Once again, 85% of the respondents indicated that, given the description of Linda, T&F was more probable than T alone.

The subjects in both studies concluded that a *conjunction* of possibilities, T&F, was *more probable* than one of the individual conjuncts, T. By the laws of probability this is not possible. T&F could, conceivably, be *as probable* as T, but not *more* so. The respondents have committed the *conjunction fallacy*. This has been taken as evidence both of people's general lack of facility with probabilities (see Chapter 2, this thesis) and of their tendency to reason heuristically; by applying the representativeness heuristic, the subjects have found the description of Linda more representative of the category "feminist bank teller" than of the category "bank teller" on its own.

Consider, however, a further demonstration of the conjunction fallacy from Tversky and Kahneman (1983: 302):

> Suppose Bjorn Borg reaches the Wimbledon finals in 1981. Please rank in order the following outcomes from most to least likely.
>
> a) Borg will win the match.
> b) Borg will lose the first set.
> c) Borg will lose the first set but win the match
> d) Borg will win the first set but lose the match.[87]

Seventy-two per-cent of respondents rated option c), "Borg loses the first set and wins the match" more likely than option b). As with the Linda vignette, the conjunction was deemed more likely than one of its conjuncts. Once again Tversky and Kahneman take this as evidence of the use of the **representativeness heuristic**. The question is, *representative of what*? Of Borg's five consecutive Wimbledon wins (1975-1980) he lost the first set three times and won the first twice. Was this the overwhelming impression of the man who dominated the final at SW19 for those years?[88]

[87] This was presented in October 1980, three months after Borg had won Wimbledon for a record fifth consecutive time,

[88] Historical note: Borg *did* reach the Wimbledon final in 1981, won the first set but lost the match in four to McEnroe.

Consider, however, that losing the first set and then coming back to win follows a traditional **story arc** in which the *hero or protagonist overcomes initial adversity to restore balance by coming out victorious at the end*. A hero who is invincible, start to finish, is no kind of story. Little wonder that subjects who have been exposed to story after story that follows such a pattern would deem this likely – or at least compelling.

Either the representativeness heuristic includes "what kind of story is this like" or we have a specific heuristic for narrative acceptability. Tversky and Kahneman sought to eliminate several kinds of probabilistic and linguistic connotations from their studies of the conjunction heuristic. It would be useful to know, given all of the thoughts examined above on the importance of narrative structures to our ordering of experience, whether we have a narrative heuristic that compares scenarios to our accepted narratives and makes judgements based on this comparison.

Defining a *narrative heuristic* would involve specifying the *process rule* the (basic and simple) capacities that the rule exploits and the kinds of problems the heuristic can solve. (Gigerenzer 2008: 24). Among the process rules would be a *search rule* such as "what kind of story is this?", a *stopping rule* akin to "stop when I find an archetype that seems to match the story I am reading/hearing/viewing" and a *decision rule* which might take the form "accept stories with elements that most closely match the archetype in content, form and emotional resonance". In the Bjorn Borg scenario (b) would be judged most acceptable (and so *likely)* a story because it matches a traditional heroic quest. The environment would be any time we are presented with information, testimony etc. from a human source. It would be a matter of empirical psychological investigation to establish which capacities the heuristic exploited.

Note that a narrative judged as acceptable, by means of a narrative heuristic, is being judged by a standard of "narrative truth" or Bruner's (1990) "verisimilitude" rather than by a truth-standard based on correspondence with the facts, or (in the case of a predicted result) against a set of calculated probabilities. In common with all heuristic judgements, its outcome is not entailed or constrained by the standards of logical inference. It is judged by whether it is a good story, whether it has meaning for us and, at base, whether it is like other stories that we have heard.

Empirical investigations – including simulations – could also establish whether such a heuristic would work and whether people use it. Given the ubiquity and power of narratives, it seems reasonable to speculate that we might.

4.9 Chapter Summary

Narratives are indispensable as a way for people to make sense of events and the stream of experience (section 4.3). They are ubiquitous, occurring in all cultures and in a variety of forms (4.1). In virtue of their central role in cognition, the narratives that we encounter have a powerful hold on us (4.3-4).

Narrative psychology (4.3) makes use of the narrative features of our mental organisation in order to construct useful pictures of the ways people process information. Narrative therapeutic approaches (4.4) use the idea that alterations to the narratives that individuals habitually apply can have beneficial effects.

Narratives, including narratives of action, are not only *possible* without allusion to the mental states of the actor, they might even be *preferable* (4.6). If the Junin Quechua people can have such rich narrative lives without terms equivalent to "belief" and "desire" and if one of the foundation texts of Western literature – Homer's Iliad – can manage to give an impression of intentional action without reference to the mental states of actors, then elimination of propositional attitudes from causal accounts need not present a major problem for our understanding of action.

Propositional attitude ascriptions are not a desirable feature of narratives because they divert attention from the action and the choices under pressure which are the traditional ways of revealing character through story (4.6). Narratives that include unnatural insight into a person's "inner life" are far removed from the way that we understand the people we encounter in everyday life. To the extent that they do appear in narratives, ascriptions of belief and desire are not core elements of the story (4.2); it is possible to render any story in which they feature without them and *still tell the same story*. They have become a part of our vocabulary of story only recently.

Once established as regular features of our common stories, propositional attitude ascriptions are accepted on the basis that they are like the other stories that we know. We use heuristic reasoning to judge whether an individual story is acceptable rather than deducing any entailed inferences about the truth of its components (4.5 & 4.8). We judge

stories by this standard of "narrative truth" (4.5) rather than by any objective standard of truth. Accepting this distinction is essential if the elimination of "belief" and "desire" from causal accounts of action is to be possible.

Narrative therapy (4.4) works by changing the stories that people habitually use to describe themselves and their situations in order to improve their psychological well-being. In this chapter I am suggesting that the "persistence of the attitudes" (Fodor 1993) owes much to the story-structure that is prevalent in our culture. Judgements of the acceptability of stories that feature propositional attitudes are based on their resemblance to other stories rather than on deductive reasoning (4.7).

Finally I speculate that because our judgement of the acceptability – or meaningfulness – of a narrative has features in common with heuristic judgements in other fields, it is conceivable that when doing so we apply a **narrative heuristic** (4.8). Empirical investigation will be required to determine whether such a heuristic is in play and what its features would be.

5 Chapter Five:

Excuses, Testimony, *Mens Rea* and Causes

Abstract: Two classes of locution are sometimes deployed to avoid embarrassment or save face, to divert moral opprobrium or to avoid legal censure: ***justifications*** *and* ***excuses****. On occasion, both of these uses are couched in terms of the beliefs, desires or other propositional attitudes of the speaker. J.L. Austin's distinction between justifications and excuses is developed here, concluding that propositional attitude ascriptions are, despite appearances to the contrary, only ever used as excuses, since justifications depend on a definition of the act itself and are therefore independent of the mind of the actor. I examine whether excuses that feature beliefs and desire are ever assessed with reference to the causal efficacy of those attitudes, together with the legal requirement of* ***mens rea****: since this concept requires that the state of mind of an accused person is considered in assessing culpability, we might expect to find mentalistic terms such as "belief" and "desire". Instead, as two case histories illustrate, a much more holistic standard of reasonableness plays a crucial role in establishing guilt.*

> In the final analysis, I can only say that I did what I believed to be the right thing.
>
> Former British Prime Minister Tony Blair, testifying to the Chilcot enquiry into the 2003 invasion of Iraq, (14th January 2011)

Frequently, we have to account to others for our actions. We have to offer justifications; we have to offer excuses. There is an important social dimension to these two functions. Whether the objective is to accept responsibility but claim that the action was justified, or to deny responsibility, the ability to avoid censure or gain the agreement of our peers is an important factor in achieving and maintaining social cohesion. There are at least three circumstances in which the need for such a description arises:

a) When the action or behaviour is unexpected, puzzling or incomprehensible given the circumstances or what is generally understood about the agent's character traits or dispositions.

b) When the action transgresses a moral norm; when a prohibition or taboo appears to have been violated or when the standard expectations of the social setting have not been met.
c) When a law is broken; when the action itself or its result is prohibited by law or when the agent fails to act in a way mandated by law in the circumstances (crimes of omission).

These are not exhaustive nor mutually exclusive. A great many violations of the law would also be abuses of moral precepts and some instances of puzzling or incomprehensible behaviour might be immoral or illegal or both. Conversely, one potential defence against a charge of having broken a law would be to claim that one was observing a moral standard that trumped the "letter of the law". One does not have to delve too deeply into imagination or history to find examples where obeying a particular law would be an example of a moral failure.

Claiming that the moral rule overrides the legal prohibition would be an example of a *justification* for the illegal act. A person who breaks some expected norm, moral or otherwise, might offer an *excuse* for their behaviour on the grounds that they were intoxicated or suffering from some other temporary defect of reason. This would not justify what they did but it would be expected to reduce their culpability or make their behaviour more understandable (in the case of divergence from expected behaviour). We even permit defences on the basis of failure of reason (although not usually in cases of intoxication) in serious legal cases.

An individual charged with an offence might use the claim that a particular action was "out of character" in *mitigation*. This would be neither a justification or an excuse.

The present chapter is concerned with justifications and excuses in one or more of these categories; in particular, with justifications and excuses that make use of claims or attributions of specific beliefs and desires. To begin, we should be clear about the meaning of "justification" and "excuse" and the distinction between them.

5.1 Justifications and Making Excuses

Austin (1979a: 176) offers an influential distinction between justification and excuse:

> In one defence, briefly, we accept responsibility but deny that it was bad: in the other, we admit that it was bad but don't accept full, or even any, responsibility.

Justifications are what we offer when we argue that the facts warranted the act. As a warranted act, either no legal or moral liability should be attached or any remaining such liability should be diminished. Neither should the act be regarded as anomalous since the circumstances justify that choice of action, however out of sorts it might appear before the justification is known.

Excuses, on the other hand, indicate that although the agent acknowledges that the act itself was unwarranted, they did not act of their own free choice – they were restricted, in other words, as to what choice of action was available to them. Examples would include coercion, involuntary movements, hypnotic states or any other circumstance in which an individual acts outside of their own rational control.[89]

If *beliefs*, for example, are fundamental to our understanding of how actions come about, then this distinction would raise two related questions, as Zimmerman (1997) points out:

> Can one do objective moral wrong and yet not act in the belief that one is doing objective moral wrong?
>
> Can one act in the belief that one is doing objective moral wrong and yet not do objective moral wrong?

In practice, as Zimmerman concedes, such questions are sidestepped. In order for self-ascriptions of specific beliefs to figure when offering excuses or justification their must, first, be some kind of *accusation*. The excuse serves the purpose of trying to deflect that accusation.

For example, a bank manager is accused of aiding a gang in robbing his own bank by opening the vault. He offers an *excuse* of coercion: he *believed* that the gang had taken his family hostage and that they would do harm to his loved ones if he failed to comply. The basis of his belief was that the lead robber had handed him a mobile telephone and he had heard his wife's voice telling him that a group of masked men with firearms were threatening her and his children at that moment. The bank manager would not deny that

[89] It is, of course, a definitional question as to whether an event that takes place outside of an agent's rational control can ever be an action at all.

opening the vault to the robbers made him complicit in their crime. He would ask that we take into account his *excuse of coercion* – made credible by his *reasonable belief* that his family would suffer harm if he did not help them. If the reality of the situation was that his wife had arranged to abscond with the gang leader and the proceeds of the robbery, the bank manager's excuse still stands on the basis of his belief in the coercion. It does not have to be real: *all that we would demand in order to find his excuse acceptable is that he had reason to believe it, reason to find the coercion compelling and that he desired to keep his family safe*. His excuse, that rests on his belief, deflects the accusation.

The question is whether this success depends on an assumption of the *causal efficacy* of his beliefs and desires?

In another example, a police officer is despatched to investigate an alleged break-in at an unoccupied house on a residential street. The call to the police has come in from a member of a neighbourhood watch association. Part of the association's service is to alert all of its members by SMS (text-message) whenever there is a suspicion that burglars might be in the area. Unfortunately, the investigating police officer misreads the address and enters a different house nearby. He makes a stealthy entry, hoping to catch the miscreant red-handed. The householder at this property is alert to the threat thanks to the message from the neighbourhood watch operation. She becomes aware of an unidentified person creeping about her house and opens fire with a handgun, shooting the unfortunate police officer dead.

Her defence, when accused of the grievous crime of killing a police officer, is that she *believed* that the police officer was a burglar. She has both the SMS from the neighbourhood watch and the police officer's method of entry and stealthy movements through her house to thank for this belief. Her defence would be one of *justification* – self-defence – had she shot the burglar. However, having in fact shot the police officer, (undoubtedly an illegal act) she is effectively, like the bank manager, offering an *excuse*, which again rests on her *belief* – in this instance that there was a burglar in her house.

Does the acceptability of her excuse rest on her belief, albeit erroneous, being the *cause* of her having opened fire? Or is our standard really one of the *reasonableness of an action under the circumstances*? Given that the same actions on the part of the householder might conceivably had led to the death of a burglar – for which she might have offered the *justification* of self-defence – where is her belief in relation to the distinction between excuse and justification?

We should not lose sight of the fact that in both of these examples we would presume that the subjects would have experienced considerable fear. The bank manager was afraid that his family was in danger, the householder that there was a burglar in her house. In fact, if in the householder case a prosecutor asked why she did not challenge the intruder by calling out, which might have given the police officer time to establish identity before the fatal shot was fired, she would be likely to cite her *fear* as an excuse. She was too afraid to make a sound.

This is reflected by the legal theorists Alexander and Kessler Ferzan (2009) in arguing for a new theory of criminal law that places the establishment of culpability as the central matter of contention. Traditional models suggest that *justifications* focus on the *wrongfulness of the act* and so should be *mind-independent* whereas *excuses centre on the blameworthiness of the actor* – and so may be supported by the actor's description of their subjective mental states, including beliefs. According to the model proposed by Alexander and Kessler Ferzan:

> ...whether the actor's conduct is justified because of the facts that actually exist or alternatively [excused] because of his beliefs about such facts will not matter because his culpability is not affected by any mistaken beliefs.
>
> (Alexander and Kessler Ferzan 2009: 92)

The motivation behind these authors' attempt to exclude discussion of an actor's beliefs from considerations of *culpability*, from the distinction between justifications and excuses and from the acceptability of either is the central question of this chapter - how much we need to regard beliefs as *causal* in order to take them into account in evaluating an excuse. This is made especially acute when other factors – fear, for example – loom large in the account. Alexander and Kessler-Ferzan's contention that we measure culpability according to the *reasonableness* of a given action under the circumstance is one to which this chapter will return.

Despite their doubt about the distinction, this chapter is principally concerned with *excuses*, because it is in the making of excuses that the beliefs and/or desires of the actor are most often expressed and considered. Justifications are concerned with the question of whether the relevant act was legitimate, given the factual circumstances under which it occurred. If an individual suggests that they *believed*, at the time of the act, that their action was justified,

then they have accepted that the act or its outcome is *not justified* and so are offering their (erroneous) belief as an *excuse*. The mental state of the actor might excuse an illegitimate act, but it cannot justify it.

5.2 The Epistemology of Testimony Applied to Excuses

Excuses and justifications are particular manifestations of **testimony**, with particular purposes. This gives rise to the question of *what we know when we are offered testimony* – the general question of the **epistemology of testimony** – and, more specifically, what we need to know in order to assess excuses couched in terms of an individual's specific beliefs and/or desires.

According to Coady (1992), testimony is a philosophically neglected source of knowledge. Kusch and Lipton (2002) count testimony among four sources of knowledge, along with perception, reasoning and memory; the distinction is that unlike testimony, these three are taken to be non-social capacities, knowable by an individual knower in isolation whereas testimony requires a testifier and a hearer. It is this, Coady claims, and Lackey (2006) concurs, that has led to the philosophical neglect of testimony as a source of knowledge. Philosophers have been prone to an "underlying assumption of epistemic individualism, according to which whatever it is possible to know, it is possible for an individual to know on her own" (Lackey 2006: 2).

Nevertheless, without an ability to give and to accept testimony, much of our life – especially social life – would be impossible. Learning is just one example of the way that we depend on the transmission of knowledge. Other information, such as travel directions, historical facts, recipes or even our own names and dates of birth come to us not from direct experience but from the testimony of others in the course of social interactions.

> We rely on the reports of others for our beliefs about the food we eat, the medicine we ingest, the products we buy, the geography of the world, discoveries in science, historical information, and many other areas that play crucial roles in both our practical and our intellectual lives.
>
> (Lackey 2006: 1)

Whether testimony should be regarded as a source of knowledge on a par with experience and reason (and if not, why not) is the central concern of the epistemology of testimony.

This field of philosophical enquiry has expanded since Coady published his book *Testimony* in 1992, possibly the first to attempt a definitive treatment of the foundations of testimonial knowledge for generations. Previous examinations – Coady cites David Hume, H. H. Price, Bertrand Russell and Thomas Reid – have principally been concerned with the sceptical paradox between our apparent reliance on testimony and our ability to judge its veracity. As well as Coady's investigations, recent work in the field has drawn upon social psychology on the nature of trust and interpersonal heuristics – for example Chaiken (1980) and M. K. Johnson et al. (1998).

As with "narrative" (Chapter 4), in spite of – perhaps because of – its ubiquity and social necessity, a definition of "testimony" seems elusive. Lackey (2006) offers this "rough characterisation":

> S testifies that *p* by making an act of communication *a* if and only if (in part) in virtue of *a*'s communicable content, (1) S reasonably intends to convey the information that *p*, or (2) *a* is reasonably taken as conveying any information that *p*.

In using the phrase "in virtue of a's communicable content" Lackey intends to exclude cases where an utterance *demonstrates* some fact rather than *testifies* to it. If, when asked "do you speak French?" an individual responds with a recitation, in the original, of their favourite passage from Baudelaire, they are not testifying, but demonstrating. The answer "yes" would be testimony, while "oui, bien sûr!" might be simultaneous testimony and demonstration. It is the affirmative *communicable content* that picks out the utterance as an instance of testimony.

Taken at face value, this would suggest that an excuse that adverts to the beliefs of the actor at the time of the act *includes testimony as to the mental state of the actor at that time*. The utterer "intends to convey" information about their belief and is "reasonably taken as conveying" that they entertained this attitude at the relevant time. We shall see in the coming discussion whether this is how such testimony is treated in, for example, the formal environment of the courtroom and in general discourse of excuse-making. Of critical importance will be consideration of the purpose to which such a statement is put.

Coady (1992) begins his examination of testimony in the legal setting and seeks to extrapolate from there to the more flexible kinds of utterance through which individuals

offer testimony in everyday discourse. Court procedures include codified rules about what is and is not acceptable as testimony. These rules are, he argues, a structured version of the more informal ways that we ordinarily judge testimony. The formal setting should thus give us a clear rule-set through which to analyse the messier styles of everyday speech.

Much of what we know is socially based and testimony is the principal method by which knowledge is transmitted from one knower to another. Coady's view, for which he acknowledges a debt to the 18th Century Scottish philosopher Thomas Reid, might be characterised as *reliabilist non-reductionism*: the hearer of testimony knows what they are told, provided the testifier is *reliable* and the propositional content of the testimony is *true*. For the recipient of testimony to acquire knowledge he need not know, or verify that the testifier is reliable; reliabilism implies that the justification for the information they require to count as knowledge is *external* to the new knower.

Austin (1979b: 82) offers a similarly non-reductive view of the epistemic value of testimony:

> The statement of an authority makes me aware of something, enables me to know something, which I shouldn't otherwise have known. It is a source of knowledge.

"An authority" in this context implies a *trusted testifier*. Since Austin is silent on how we are to judge whether a particular testifier is to be trusted, we should perhaps take him to be *externalist* about the justification of knowledge acquired in this way. It is a sound description of the way that we treat everyday testimony. If someone that we trust tells us something and we have no reason to doubt what we are told – no **defeater**, in the language of epistemology – then we tend to regard to content of the testimony as new knowledge.

The reductionist view of testimonial knowledge suggests that in evaluating testimony we are compelled to take account of other, non-testimonial background features, both to assess the degree of trust to place in the testimony (the likelihood that the content of the testimony is true) and the degree of trust that we have in the testifier (has this person ever lied to us before? What would they have to gain by lying about this? Could they be mistaken? Reductionists suggest that "knowing because we have been told" is always reducible to and dependent upon non-testimonial knowledge; even if that knowledge amounts to a simple rule like "people [that person] usually tell[s] the truth, unless they have a reason to lie".

The distinction is further developed by Pritchard (2004) who writes that reductionists "argue, roughly, that the justification of an agent's [testimony based belief] is always dependent upon that agent possessing further independent grounds – i.e., at the very least, grounds that are independent of the instance of testimony in question". Non-reductionists like Coady – Pritchard dubs them *defaultists* or *credulists* – on the other hand, maintain that "the epistemic status of a testimony-based belief need not depend upon the agent possessing any independent grounds in favour of that belief. *Just so long as there are no grounds for doubt* …" (emphasis added).

Lackey (2006) subdivides the reductionist view into *global* and *local*. The global reductionist takes *all* testimony-based knowledge to be dependent on positive, non-testimonial reasons for accepting a report. The local reductionist demands that "in order to justifiably accept a speaker's testimony, a hearer must have non-testimonially based positive reasons for accepting the particular report in question.

Bearing in mind the overwhelming role that testimony plays in our epistemic lives perhaps we *ought* to default to acceptance – in fact perhaps, in accord with Austin's suggestion above, we *do* default in that direction. This is the basis of the "acceptance principle" suggested by Burge (1993: 467):

> A person is entitled to accept as true something that is presented as true and that is intelligible to him unless there are stronger reasons not to do so.

This normative statement could be a manifesto for non-reductionism, in which testimony is to be accepted unless there is a *defeater* for acceptance – a reason to doubt.

In the case of excuses, however, there is *always reason to doubt*. Those reasons lie in the purposes to which people put justifications or excuses which always entail avoiding censure, deflecting criticism or saving face – purposes over and above that of getting the hearer to accept the testimonial content of their utterance or to relay knowledge. We should analyse a typical use of claims of specific beliefs and desires in offering excuses against the various theories of the philosophy of testimony to determine whether:

a) The reason-giver's specific belief claims are *offered* as testimony.
b) The reason-giver's specific belief claims are *considered* as testimony.
c) The reason-giver's specific belief claims are offered or considered as *causal* claims.
d) What other purposes they might serve.

We might think of the transmission of testimonial knowledge as like a "bucket brigade" (Lackey 2006: 6) whereby full buckets of water are passed along a chain of people to reach a fire. Each person in the chain can pass a full bucket to the next only if they have a full bucket to begin with. In order for a hearer to be passed knowledge in the form of testimony, the speaker must have that knowledge to begin with.

On occasion we do include ascriptions of specific beliefs and desire when seeking to excuse a third party's anomalous behaviour. According to the bucket-brigade simile, a person hearing this only gains knowledge of that third party's beliefs and desires if the person who tells them has *knowledge* of those mental states. However, their understanding of the mental states of a third party must be limited to:

a) The third-party's verbal statement of what they believed and desired.
b) Inferences based on something like the belief-desire law.

Both of these sources have limited value as a source of testimonial knowledge: a) because first-person excuses are illocutionary acts which always perform functions other than making a claim about the actor's mental states (see below) and b) because such inferences are only ever "as if" explanations; any number of potential belief-desire pairs might fit any particular action according to the belief-desire law.

Some impressions of a third-party's performance of an act might be salient. For example, some features of the performance – how rapidly it was performed and with how much care, for example, might lead the observer to make inferences about how *deliberately* it was done. Acting deliberately, according to Austin (1979c: 286) implies deliberation, or a weighing of the alternatives and a choice between them – including the option to refrain from committing the act at all. Deliberation requires time and the opportunity to reflect. An observer might also take into account the agent's actions leading up to the event, for indications of pre-meditation or planning. The circumstances will always play apart (situational attribution); was the lighting good, for example? Was the agent in an unfamiliar place? Were they familiar with the customs or routines of any other participants?

These inferences are always subject to revision in the light of new information – such as when we *ask an agent whether they acted on purpose*. We might also learn more about the individual to make sense of their behaviour. If we witness a person lash out in panic at the sudden appearance of a dark shadow on a wall, we might consider them a nervous type –

until we learn that they had recently been the victim of a violent assault. In this light we revise our opinion and find the behaviour more understandable.

When confronted with behaviour that is criminal, morally reprehensible or anomalous we typically seek reasons for the action. We would ask questions. The answers would not usually include allusions to specific beliefs and desires.

> Q. Why did Sheila take up ballroom dancing?
> A. She *desired* to improve her fitness and *believed* that dancing would be an enjoyable way to get fit.
> Q. Why did Gary kick the dog?
> A: He *desired* to avoid suffering an allergic reaction and *believed* that a kick would discourage the animal from bothering him.
> Q. Why did Louie murder Stan?
> A. He *believed* that Stan was an informer and *desired* to cut off the flow of information to the authorities and to discourage others.

These are all unnecessarily awkward constructions. In the first two examples more usual responses would be simply "To get fit" or "He's allergic to dogs". It might be argued that the propositional attitudes are tacit, or assumed; *but the specific thought processes of the agents are not directly alluded to and might conceivably be different* – Sheila might take up ballroom dancing because she fancies the instructor and Gary might kick the dog because, perhaps as a result of his allergy, he hates dogs. Only in the third example is a specific epistemic condition implied – Louie thought that Stan was an informer. In each case, the usual answer is in the form that we might expect if the question was couched as:

> Q: If I asked Sheila why she took up ballroom dancing *what would she say*?
> A: That it's a fun way to get fit.
> Q: If I asked Gary why he kicked the dog, *what would he say*?
> A: That he's allergic to dogs.
> Q: If I asked Louie why he killed Stan *what would he say*?
> A: That Stan was an informer.

Very often, third party accounts like this are proxies for asking the agent themselves. They are the observer's impression of what *reason* the agent would give if interrogated. Their actual motivations, in all their complexity, are unavailable.

This is why we exclude from court proceedings testimony about the beliefs or other propositional attitudes of third parties. A witness can testify about the actions that they witnessed, and even about the outward demeanour, facial expressions, gestures and speech that they saw and heard: speculation – or even deductions – about the mental states of the agent at the time the act was performed is not admissible as evidence. Like the person in the bucket brigade, legal testimony is limited by procedure to what the witness has to pass on – not hearsay, not speculation, but those things that they can legitimately claim to *know*.

This brings out an important distinction between third party accounts of anomalous, immoral or unlawful behaviour and *excuses* that are offered by a person accused of such an action. The offering of an excuse has an illocutionary function other than explaining or describing the antecedents of the action. The person offering the excuse is always in some jeopardy. They might risk being socially ostracised for their "strange behaviour", condemnation and distrust for their immorality or the loss of their liberty or even life for their illegality. The principal use of the excuse is to minimise the possibility or severity of these potential consequences. It is to this end that, on occasion, self-attributions of specific beliefs and desires sometimes feature.

5.3 Are Belief-Desire Excuses Intended or Evaluated as *Testimony*?

If a housemate takes and eats the last slice of a cake from the communal refrigerator in a shared house they might, when challenged offer the excuse that they thought (believed) that this slice was intended for them. They have acknowledged that the cake was not, in fact, their own but are offering their erroneous belief as an excuse. There are several elements to this statement, some of which are testimony and some of which are not.

> Statement: "Yes, I ate the last slice of cake. But I thought that it was meant for me!"

The first sentence is both testimony to the fact and an *acceptance* of the accusation, a recognition that the accusation has a factual basis.

The second sentence is an excuse. As well as offering the utterer's *reason* for having eaten the last slice, it is suggesting that the owner of the cake *ought* to have left a slice for them. As well as a reason, it is an attempt at a partial justification – suggesting that since the cake *should* have been shared, in taking it they were doing no more than following that *norm,* at least in their estimation.

Recall that the first test of testimony contained in Lackey's "rough characterisation", testifying to a proposition (p) requires that "in virtue of [the statement's] communicable content" the utterer reasonably intends to convey the information, p.

The cake-eater is not testifying to the possession of a particular belief about the provenance of the sweetmeat. They are deflecting opprobrium and denying that they are the sort of person that would steal and eat somebody else's cake under normal circumstances. They are less concerned that the information coded in "I thought that it was meant for me" is taken to be true than that they should avoid censure and that they should not be regarded as the kind of person to help themselves to someone else's cake. This implies that the portion of their statement including an alleged description of their beliefs fails the first test to qualify as testimony. By stating this they do not intend, in virtue of the statement's content, to convey information about their thoughts at the time the cake was eaten.

The second possibility (Lackey offers these two either side of an "or" conjunction) is that the statement of p qualifies as testimony if it is "reasonably taken as conveying any information that p". Is an excuse that features the erstwhile belief of the excuse-giver reasonably taken as conveying information about their beliefs?

When the householder in the earlier example above finds herself in a court of law, charged with the unlawful killing of a police officer, she would offer as an *excuse* that she thought (believed), when she pointed a lethal weapon at a human target and fired, that her life was in danger from the presence of a burglar, that her action in deploying deadly force was *justified* by the imperative that she protect herself. Although she now recognises that the self-defence justification is not available since the police officer she killed presented no danger, she asks the court to accept her excuse because:

a) She *believed* that her life was in danger.
b) This belief was reasonable in the circumstances.
c) Any reasonable person might have done the same.

The third of these is highly significant. This would be the test that any court would apply – was the action of the accused *reasonable in the circumstances*? The court will not concern itself with the attribution of specific states of belief; it is just as likely that the householder, having been warned that a burglar was in the area and confronting a figure sneaking about in her darkened house was too terrified to think rationally or to entertain such detailed

thoughts as "I believe that person is a burglar and desire to rid myself of the threat that they represent and further believe that discharging my firearm is the best way to bring this about". Panic is a better explanation of why she did not verbally challenge the police officer, or call the police, or even switch on a light.

The information that she received from the neighbourhood watch is certainly pertinent to the reasonableness of her actions. This might be construed as a support to her claim of a specific belief. However, the attribution of a specific belief at the time she pulled the trigger is not the test that the court will apply. More significant is the question of reasonableness. In the same circumstances, including (but not limited to) the receipt of the text message from the neighbourhood watch, were her actions consistent with those of a reasonable person? If they are, her excuse stands; if not, it fails.

Consider if another householder's reaction to the neighbourhood watch text message had been to run into the road shooting wildly and maiming one of his neighbours. If his excuse was that he believed that the neighbour was a burglar, we would not accept it – not because we question his belief but because we question the reasonableness of his actions. And yet he had the same information, the same justification for his claim of belief as did the first householder.

The belief-component of the excuse is a part of establishing reasonableness. As such in the case of an excuse, the claim of a specific belief is not assessed as testimony as to the belief of the person giving the excuse and as such the excuse-maker is not offering testimony as to the content of their belief.

5.4 Are Excuses Assessed as *Causes*?

In certain circumstances the belief and desire component of an excuse is accepted as evidence of the individual's degree of culpability. This has the superficial form of a causal claim – that it was the beliefs and/or desires that the agent entertained at the time of the act that *caused* them to act in the way that is being considered. The question that arises from this is whether accepting or rejecting an excuse in those terms involves the assessment of those specific mental states as *causes.*

There are a number of suggestions as to how we distinguish the causes of an event from other correspondences. For brevity, let us consider one – the idea of *counterfactual*

causation. Kim (1973) characterises the account of counterfactual causation that he finds in (Lewis 1973) in these terms:

> (1) An event (e) causally depends on an event (c) just in case if (c) had not occurred (e) would not have occurred.
>
> (2) An event (c) is a cause of an event (e) just in case there is a chain of events from (c) to (e), each event in this chain being causally dependent on its predecessor.

Kim points out that this also entails that the counterfactual conditional "If (c) had not occurred then (e) would not have occurred" describes a situation (on Lewis' reading) in which (c) causes (e).

For the purposes of this thesis, I can leave aside the metaphysical questions of causation and the application of Lewis' *possible world semantics* to the consideration of counterfactual reasoning. The question at issue is whether we assess excuses couched in belief-desire terms against a counterfactual standard of causal reasoning. The specific question regarding excuses is:

> Q1: In judging the acceptability of an excuse couched in belief/desire terms do people apply counterfactual causal reasoning or apply some other standard?

We can formulate the requirements of the belief-desire law into a counterfactual test of an ascription of a specific belief-desire set:

> A specific intentional action, performed by an actor, is caused by that actor's specific beliefs and desires at the appropriate time,[90] iff had that actor not possessed those specific beliefs and desires at that time then they would not have performed that action.
>
> An actor's specific beliefs and desires at the appropriate time is the cause of their specific intentional action iff there is a chain of events from that belief-desire set to the action, each event in that chain being causally dependent on its predecessor.

[90] I am aware that "at the appropriate time" is question begging, since the "appropriate time" to cause the behaviour presumes that the behaviour is so caused. However, I will leave this consideration to one side for the present.

Excuses couched in belief-desire terms might take these forms:

a) I did not *believe* that acting as I did would bring about the culpable result that in fact occurred.

b) My *desire* to bring about a good (non-culpable) result inadvertently resulted in the culpable action that I am accused of

c) I *believed* that I was (legally, morally, etcetera) obliged to act in the way that I did.

d) Given what I *believed* about the situation, any reasonable person would have acted as I did.

All of these imply error, although of different kinds. Excuses of the kind a) comprise a failure to foresee the consequences of the action. The second kind, b) is also a failure to see the consequences but is principally a claim as to the *motivation* of the action, an assertion that the individual's *intentions* were good. The third, c) suggests an error as to the expectations under which the individual was acting; again the utterer is claiming good intentions – compliance with a norm or a law – but admitting that they might have been mistaken as to their obligations. The fourth, d) suggests that although the *information* (content of their belief) might have been wrong or inadequate, had any "reasonable person" been similarly informed, then they would have acted similarly.

The counterfactual test of all of these would be to ask:

> Had this accused person, or any reasonable individual, *not* possessed the precise belief or desire described in the excuse, would they have behaved differently?

If one accepts both the causal efficacy of specific beliefs and desires and the application of counterfactual causal reasoning to the assessment of an excuse that makes use of those terms, then we should expect the *rejection of an excuse* to take one of these forms:

i) The actor did not entertain the precise belief and/or desire specified in their excuse at the appropriate time.

ii) The relationship between the specific belief and/or desire claimed in the excuse and the action taken is not plausible.

iii) Had the actor possessed a different specific belief and/or desire to those specified in the excuse, then they would not have acted substantially differently. (that is, the claimed beliefs and desires offered as an excuse made no difference to the outcome.

In practice, I suggest, none of these is the standard against which we usually assess such reasons. In the case of a) for example, the starting point of such an assessment is concerned *not with whether the action was caused by the belief but with whether the person should have had that belief at al*l. At b) the question at issue is not whether or not had the accused person not had that desire (to do good, or to avoid a potentially worse outcome) would they have acted the same way but *whether it is conceivable that a well-motivated person could have carried out that action, regardless of the specific desire or its presumed causal relationship to the action*. When judging c), our principal concern is not with the relationship between the belief and the action but with *whether the belief claimed was reasonable*. It might be claimed that this presumes a causal relationship between the belief and the act, but we do not determine the acceptability of the excuse on the basis of whether the person would have acted differently had they entertained a different belief.

The final example, d), might also be taken to suggest that had the information, or belief, been different then the actor would have taken a different course. However, our assessment of the acceptability of this kind of excuse does not turn on this relationship but on what is meant by reasonable. We apply a standard of reasonability to actions not on the basis of any relationship between beliefs or desires and an action but on the basis of the totality of the circumstances, the features of the actor and, crucially, on the degree of risk attached to the circumstances of actions. This, in turn, directs us to accept or reject an excuse based on *what is at stake*. Our condemnation of the person offering the kind of excuse at d) will depend on whether we think that they should have taken more care – whether they were being reasonable in the situation – rather than on whether people would *always be caused to act that way* given the same content of their beliefs. This distinction, the notions of risk and reasonableness, will be developed further in what follows.

One reason to reject the automatic acceptance of excuses as a source of testimonial knowledge is that we always have a *defeater* for the truth of an excuse – the giver of an excuse has other motives for offering it. These are the socially significant functions of diverting blame or the opprobrium of peers. In the case of excuses given in the

formal environment of a courtroom, the jeopardy faced by a defendant might be even more pressing. If held fully accountable for the action under examination, such an individual faces the possibility of being fined, a loss of liberty or even, in some cases in some jurisdictions, a loss of life. We can ask another question:

> When someone is accused of some anomalous, reprehensible or culpable behaviour and offers an excuse that refers to specific beliefs and desires that they say they possessed at the time of the act, *would they offer the same beliefs and desires to explain their behaviour were they not in jeopardy*?

This is a different kind of counterfactual question, enquiring not about the causal efficacy of specific beliefs and desires but about *why people might offer them in excuses*. For any action, there might be any number of reasons. Some of these will advert to situational attributions as suggested by attribution theories in social psychology. Some, exhibiting the *fundamental attribution error* (Chapter 3, section 3.4) might call on the stable trait characteristics of the actor even though situational factors are sufficient to explain it. Most of the time we leave beliefs and desires unspoken. However, when there is anomalous behaviour to be explained – or when we need an excuse – we sometimes offer these kinds of ascriptions.

This goes beyond the fundamental attribution error into *unstable personal attributions*. If our expectation is that excuses will provide additional *information* about the *causes* of a particular action, then we have lost sight of the purpose for which reasons are offered. When John Dillinger, an infamous outlaw of the US in the 1930s, was asked why he robbed banks he answered "that's where they keep the money." This is a resolutely situational attribution. Dillinger was also offering no excuses. Neither did he suggest additional *causes* of his choices.

If the causes of an action are adequately understood *before* an excuse is given *and* we must judge an excuse adverting to beliefs, desires or other propositional attitudes by some causal standard, then we are in danger of admitting an **overdetermination of causes**. Why should it be that when a person in jeopardy offers an excuse we now have additional causes to consider? Bear in mind that the original causes have not gone away – they are still salient to our understanding of the event. We still recognise that what happened was caused by the intentional actions of the agent. When we invite an excuse by asking the question "why did you do that?" we are not looking for more causal-explanatory information but for better

understanding of the individual, for them to tell us *why we should not condemn them*. Superficially we may be confused by the use of "why" in the question, in that it looks similar to an appeal for a causal explanation. This was a source of confusion of which Wittgenstein was well aware.

> The double use of the word "why", asking for a cause and asking for the motive, together with the idea that we can know, and not only conjecture, our motives, gives rise to the confusion that a motive is a cause of which we are immediately aware, a cause 'seen from the inside', or a cause experienced.
>
> Wittgenstein, *The Blue Book*, (1958: 15)

5.5 Guilty Acts and Guilty Minds: *Mens Rea* and the Law

Under many of the laws administered within common-law[91] jurisdictions there is an explicit reference to the intentional states of the defendant. This is the notion of ***mens rea***. In the case of certain common law offences, and even some statutory requirements, the presence or absence of *mens rea* is a determinant not just of the guilt of the defendant but also of *whether the offence has, in fact, been committed*. The clearest example is the common law conception of *murder*. The principal difference between what was once known as "wilful" murder and lesser offences of unlawful killing is the *mens rea* of the defendant. Murder is the killing of a person where the *intention* is to kill or to cause serious harm. *Absent the intention, the crime of murder has not been committed*.

The phrase in mediaeval Latin that captures the intentional aspect of this requirement is *"actus non facit reum nisi mens sit rea"* (Sayre 1932) which translates as "the act is not made guilty unless the mind is guilty'. Culpability for an offence with a *mens rea* requirement has two components:

i) The event which transgresses the legal prohibition – the *actus reus*.
ii) The concurrent intentional state of mind that renders the transgressor culpable – the *mens rea*.

[91] Throughout the discussion of legal procedure, I am taking the example of those jurisdictions that base their approach on the Common Law. This includes the United Kingdom, most of the Commonwealth including India, the United States (except Louisiana) and so includes a considerable proportion of the global population.

Philosopher of law Michael S. Moore points out that this distinction is less clear cut when one considers that the identification of any event as an *act* implies that it is intentional, that the actor must will his own movements. That is not, however, sufficient for the *mens rea* requirement:

> …the intention required to act at all – the intention to move one's limbs – is not the same in its object as the intention described in the *mens rea* requirement. The latter intention has as its object complex act descriptions like 'killing', 'disfiguring' or 'recording a confidential communication'; it does not have basic act descriptions like 'moving my finger' as its object.
>
> (Moore 1993: 173)

This excludes simple accidents. If an engineer reaches for a tool and catches his sleeve on the handle of his toolbox, causing it to fall into the aero engine on which he is working, the act of moving his arm was clearly intentional but the costly damage to the engine was not. He is not responsible in the same way that would be had he *thrown* the toolbox into the engine, causing similar damage.

Mens Rea has two categories. **General *mens rea*** is the requirement that the transgressive outcome of the action must not be inadvertent or accidental – such as the the engineer inadvertently causing his toolbox to fall into the engine. General *mens rea* gives *insanity* or *diminished responsibility* defences their traction: even though an act and its outcome might be *willed* certain individuals are considered to be lacking in control to the extent that they are deemed incapable of choosing to act differently (Morse 1999; Phillips and Woodman 2008).

Specific *mens rea* is that component of certain crimes that specifies an intent to commit precisely the prohibited act as part of the *definition* of the offence. The most famous example of specific *mens rea* is the *intent to kill* component of murder.

Sayre (1932) wrote that "No problem of criminal law is of more fundamental importance or has proved more baffling through the centuries than the determination of the precise mental element or *mens rea* necessary for crime." Almost sixty years later, according to Morse (1991) the situation had not been clarified: "Few legal terms confuse behavioural scientists and mental health professionals more than *mens rea* (guilty mind), largely because the law employs the term in diverse and often inconsistent ways."

This confusion prompted the American Law Institute to attempt a clarification of the term "*mens rea*" in their Model Penal Code. For the purposes of framing the *mens rea* component of an offence the following key terms are defined:

> **Purposely**. A person acts purposely with respect to a material element of an offence when: (i) if the element involves the nature of his conduct result thereof, it is his conscious object to engage in conduct of that nature or to cause such a result; and (ii) if the element involves that attended circumstances, he is aware of the existence of such circumstances or he *believes or hopes* that they exist.
>
> **Knowingly**. A person acts knowingly with respect to a material element of an offence when: (i) if the element involves the nature of his conduct or the attendant circumstances, he is aware that his conduct is of that nature or that such circumstances exist; and (ii) if the element involves a result of his conduct, he is aware that it is practically certain that his conduct will cause such a result.
>
> **Recklessly**. A person acts recklessly with respect to the material element on offence when he consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that, considering the nature and purpose of the actor's conduct and the circumstances known to him, its disregard involves a gross deviation from the standard of conduct that a law-abiding person would observe in the actor's situation.
>
> **Negligently**. A person acts negligently with respect to a material element of an offence when he should be aware of the substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of the actor's conduct and the circumstances known to him, its disregard involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation.
>
> (American-Law-Institute 1985 Sec. 2.02)
> (Emphasis added)

According to Alexander and Kessler Ferzan (2009: 23), this model attempts to reduce the "proliferation" of "mental state concepts" in the *mens rea* components of offences in various

jurisdictions to just these four. Morse (1991: 212) and Yeager (2006) independently point out that these are presented in descending order of culpability. At all levels, the nature of the *risk* determines the degree to which a transgressor should be regarded as culpable. This echoes Austin's observation that:

> The extent of the supervision we exercise over the execution of any act can never be quite unlimited, and usually is expected to fall within fairly definite limits ('due care and attention') in the case of acts of some general kind, though of course we set very different limits in different cases. We may plead that we trod on the snail inadvertently: but not on a baby – you ought to look where you are putting your great feet.
>
> (Austin 1979a: 194)

The usual way in which we assess culpability is three fold: the degree to which the action under consideration was intentional, the degree to which it was the accused person's *intention* to cause the harm and an assessment of what is at stake. In Austin's example, a person who inadvertently steps on a snail might have the same thought processes as one who steps on a baby: where a baby, rather than a snail is at risk, we demand a higher standard of attention to the risk (Yeager 2006: 60).

An understanding of this distinction prompts legal theorists Alexander and Kessler Ferzan to propose an interpretation of criminal law under which culpability is assessed according to an *appreciation of risk* rather than more vague and less easily determined measures such as *responsibility* or *intention.*[92] "an actor's culpability consists in his imposing a risk to others' legally protected interests for reasons that do not justify imposing the risk (as he assesses it)" (Alexander and Kessler Ferzan 2009: 86). Their approach entails that the role of the court is to assess a defendant's offer of a justification or an excuse against a standard of the *reasonableness* of their *assessment of risk*, rather than any evaluation of the truth of any ascriptions of specific beliefs or desires that might accompany the justification or excuse. They write:

[92] For much the same reason that an ascription of specific beliefs and desires is explanatorily inadequate, using these notions to assess culpability implies an unattainable degree of access to an agent's thought processes.

> Whenever the actor's reasons are sufficient to justify the risk, the actor is justified. Even if the actor's actions were not justified, if she has lived up to all that we can really expect of her, then she is excused.
>
> (Ibid.: 87)

As well as distinguishing accidental or incidental events from intentional actions, an important additional function of *mens rea* components in the drafting of criminal law is to leave open *mens rea* defences (Morse 1991). If a defendant can establish that, at the time the alleged offence was committed (or in the moments immediately prior to its commission) they did not have the requisite guilty mind, then they have a defence, or at least a partial defence. This defence is based on the failure to prove, to the required standard ("beyond a reasonable doubt", for example) that the defendant possessed *mens rea*. It is not a requirement that they prove possession of alternative thought processes at the time: only that their circumstances, behaviour, understanding, and so on *raise reasonable doubt* that their thinking constituted *mens rea*.

In such a circumstance, a defendant might wish to offer a defence based on the self-ascription of specific beliefs and desires coinciding with the *actus reus*. It is not for the defendant to prove that this alleged combination of states was *true* or *causally efficacious*: it is sufficient for their defence to succeed if their suggestion of these specific propositional attitudes introduces reasonable doubt of their *mens rea*. The adversarial nature of criminal proceedings in common law jurisdictions seeks to establish whether any doubt so engendered is *reasonable*, in the view of the judge or the members of the jury.

The task for the prosecution in a common law criminal trial is to establish, beyond a reasonable doubt, that the defendant not only physically transgressed the bounds of the law – that there is an *actus reus* to be answered[93] – but also that they did so deliberately, wilfully, volitionally or, in some cases, maliciously (Edwards 1955). *Mens rea* cannot be established forensically – thoughts leave no fingerprints – or by any other fact-based enquiry. Judges and juries must infer the *mens rea* on the basis of cues to do with the salient features of the act (including the degree of risk), the defendant's behaviour before, during and after the commission of the act and reports of their demeanour at the same times. Some features of the act may also indicate how *deliberately* the defendant acted. It would be reasonable to

[93] Defences of alibi, or disputes as to the facts of the defendant's actions take precedence over mens rea defences.

presume that a defendant who used a hypodermic to lace a cream cake with weedkiller, then offered it to their victim over tea, has shown sufficient planning to fulfil the *mens rea* requirement for a charge of murder – what used to be known in English law as *malice aforethought*.

This antique phrase does capture something of the specific *mens rea* component of this most serious of crimes. For a murder (rather than manslaughter or some other class of unlawful killing) to have been committed, it is not sufficient that the defendant intended the act that led to the death of another. They must intend harm – to the actual victim or to some other person (as in cases of mistaken identity, or if someone other than the intended target eats the poisoned cake). For example, if somebody steals the cabling at the side of a railway line and, as a direct result, a motorist is killed on a level crossing some miles from the incident, the thief is unlikely to be accused of murder. The theft was a deliberate act, but there was no direct intention to kill or to cause harm to persons.

In some cases, facts about the offence are sufficient to establish *mens rea*. If a burglar enters a property and is tackled by the householder, who is killed by a bullet from the burglar's gun, a court is unlikely to accept a *mens rea* defence on the grounds that the burglar intended only theft, that the discharge of the firearm was accidental. By taking a weapon into the commission of a crime the burglar has exhibited *mens rea* to the extent that they came prepared to use deadly force to resist capture (Carson and Felthous 2003). If the gun belonged to the householder, however, a *mens rea* defence might be available – so long as the burglar had not gained full control of the weapon.

Despite the assertion in Morse (2007 emphasis added) that "The law's view of the person is a creature capable of *practical reason*, an agent who forms and acts on *intentions that are the product of the person's desires and beliefs*", which is a rehearsal of the view from philosophical folk psychology, ascriptions of specific beliefs and desire do not figure in the *mens rea* components of prohibited acts, or of unlawful omissions. According to Edwards (1955) the *mens rea* components of statutory offences include adverbs such as "maliciously", "wilfully", "knowingly" or "fraudulently" or verb-forms that imply responsibility, such as "causing", "permitting", "suffering" or "allowing". "Belief" and "desire" might feature in a defendant's attempt at exculpation, but not in the *mens rea* requirements of the relevant law.

The *mens rea* components in the drafting or understanding of laws exist to excuse accidents and mistakes (Morse 2003) to distinguish wrongful from harmful (Yeager 2006: 86) and to allow room for defences in which the intentions of the accused serve to justify or excuse the action (Morse 1991). Conceivably, excuses might involve self-ascriptions of specific beliefs and desires; however, these are assessed for whether they introduce reasonable doubt as to the defendant's *mens rea*, rather than as claims of the causes of their actions.

In support of the claim that the law expressly treats the thoughts of an accused in the way that I have suggested, we should consider two rather different cases, each turning on the *mens rea* of the defendants.

5.6 Case Study: R v Morgan et. al (1975)

In 1975, three men were convicted of rape and, along with a fourth, the eponymous Morgan, of being a party to rape,[94] The circumstances of the incident were that Morgan, an RAF officer, after a night of heavy drinking, invited three companions to his home where the three were to engage in sexual activity with his wife while Morgan watched. When charged with rape, the three based their defence on an assertion that Morgan had told them that his wife would enjoy the experience and that her resistance and protestations were merely her way of enhancing her enjoyment. As a result, they claimed, they had no *mens rea* for rape. They had *believed* that Mrs Morgan had consented and that the act, therefore, did not constitute rape. The three maintained this defence in spite of the fact that the victim had suffered considerable physical injury and had cried out to her 11-year-old son, who was asleep in the next room, to call the police (Curley 1976).

The three men were convicted of rape and lost their appeal on the grounds that the jury was deemed right not to have accepted their testimony. However, the appeal court justices noted that the case had raised a significant issue over the *mens rea* component of the crime of rape. As the law was then formulated, the crime of rape included a specific *mens rea* component that the accused *had to intend to engage in sexual intercourse with a person against that person's wishes*. This would entail that an "honest belief" that the alleged victim had consented would be a defence, in that it would exclude the specific *mens rea* of an intent to rape.

[94] The legal reference for this case: Sub. Nom. R. v. Morgan [1975] 1 all ER 8

At the original trial the judge had directed the jury that:

> …his belief must be a reasonable one; such a belief as a reasonable man would entertain if he applied his mind and thought about the matter. It is not enough for a defendant to rely upon a belief, even though it be honestly held, if it was completely fanciful; contrary to every indication which could be given to carry some weight with a reasonable man.
>
> (Curley 1976)

The problem was that the standard of *reasonableness* – that a "person of reasonable firmness" (Morse 2007) would act the same way – was not, at that time, enshrined in the law. For this reason, the three men had based their appeals on a contention that the trial judge, in so directing the jury, was in error. Although rejecting the appeal on the ground that the jury would have found the three guilty even in the absence of this direction, the Court of Appeal allowed the case to go on to the House of Lords[95] so that this anomaly could be addressed at the highest level.

The House of Lords also rejected the appeal because, they ruled, a properly directed jury would have found the men guilty –their testimony as to what they had *believed* at the time of the act would have been rejected. In their judgement, however, the Law Lords conceded that the original trial judge *had* misdirected the jury as to the specific *mens rea* component of the crime of rape. One of the sitting judges, Lord Cross, wrote in his judgement that:

> ... the question to be answered ... is whether according to the ordinary use of the English language a man can be said to have committed rape if he believed the woman was consenting to intercourse and would not have attempted it but for his belief, whatever his grounds for so believing. I do not think that he can.

Another, Lord Hailsham, set out his concern with the law:

> …the prohibited act in rape is non-consensual sexual intercourse, … the guilty state of mind is an intention to commit it, it seems to me to follow as a matter of inexorable logic that there is no room either for a "defence" of honest belief or mistake, or of a defence of honest and reasonable belief and mistake. Either

[95] The highest UK court, at the time.

> the prosecution proves its case or it does not. Either the prosecution proves that the accused had the requisite intent, or it does not. In the former case it succeeds, and in the latter it fails. Since honest belief clearly negatives intent, the reasonableness or otherwise of that belief can only be evidence for or against the view that the intent was actually held ...

This signifies that the law would allow as a defence against rape the defendant's belief that the other party had consented, *whatever procedure the accused had used to arrive at that belief*. The only grounds for rejecting that defence and convicting would be the rejection of the claim of the specific belief. This would make rape convictions even harder to achieve *because the onus would be on the prosecution to prove that the defendant did not believe what they claimed to believe*. If a jury had reasonable doubt as to whether the prosecution had proved this, then they should acquit.

Would it ever be possible for a prosecution to prove, beyond a reasonable doubt, that a defendant *did not believe what they claim to have believed*?[96]

In response to this controversy, the UK parliament rushed through the Sexual Offences (amendment) Act, 1976. In Section 1, part 1 this provides that:

> (1) For the purposes of section 1 of the Sexual Offences Act 1956 (which relates to rape) a man commits rape if -
>
> (a) he has unlawful sexual intercourse with a woman who at the time of the intercourse does not consent to it; and
>
> (b) at the time he knows that she does not consent to the intercourse or he is reckless as to whether she consents to it;

The inclusion of the standard of recklessness is allows the jury to decide whether or not a *mens rea* defence against rape is *reasonable*. The prosecution does not have to prove that the defendant did not believe that consent had been given, only that his behaviour indicated a recklessness as to whether the victim had consented.

Almost thirty years later the specific *mens rea* component of rape was further clarified under the Sexual Offences Act (2003), which at Chapter 42, Part 1 defines rape:

[96] Which is a central problem with the idea that actions can be explained or predicted by the ascription of specific propositional attitudes.

(1) A person (A) commits an offence if—

(a) he intentionally penetrates the vagina, anus or mouth of another person (B) with his penis,

(b) B does not consent to the penetration, and

(c) A does not reasonably believe that B consents.

(2) Whether a belief is reasonable is to be determined having regard to all the circumstances, including any steps A has taken to ascertain whether B consents.

This allows room for the *mens rea* defence of "I believed that the other party had consented", but ensures that the belief is judged not on an impossible standard of truth or falsity, nor of causal efficacy but *against a standard of reasonableness*. The best way to ensure that someone has consented is to ask the question: in the absence of that polite enquiry a jury would be justified in rejecting their subsequent defence on the grounds that they *believed* that they had consent. The action has to be *reasonable in the circumstances*, including asking for consent: considerations of whether their belief is actual, honestly held or causally efficacious do not enter into the deliberation.

5.7 Case Study 2: "Let him have it Chris!"

The notorious killing of Police Constable Sidney Miles in November 1952 gave rise to what many regard as a grave miscarriage of justice (Yallop 1971).[97]

On the evening of Sunday 2nd November 1952 two teenaged petty criminals, Christopher Craig (aged 16) and Derek Bentley (19), having failed to break into a confectionary warehouse in Croydon, found themselves on the roof of the building. Neighbours had already called the police after seeing the pair climb over the fence into the premises. Within minutes of the call a detective constable Frederick Fairfax had arrived accompanied by two uniformed officers, constables Pain and Bugden. Almost simultaneously two more uniformed policemen arrived, constables McDonald and Sidney Miles.

The two would-be burglars were trapped on the seven-metre high roof, unable to escape the way they had come without running straight into the police and immediate arrest. They hid behind the head of a lift shaft, which was the larger of two features on the otherwise flat

[97] I have relied on Yallop's detailed account for much of the factual detail here. I have avoided much of the commentary Yallop offers as this was directed at highlighting the perceived wrong done to the defendant, Bentley.

roof. DC Fairfax climbed to the roof the same way that Bentley and Craig had and called to them to come out from their hiding place. Bentley emerged and Fairfax took hold of him. As Fairfax tried to round the lift head, still holding Bentley, Craig stepped out, produced a revolver and fired. The bullet struck Fairfax superficially in the shoulder and he fell to the ground. Bentley, the detective would later testify, pulled back the shoulder of his coat and asked "You all right?". The shocked officer regained his composure and ran to the edge of the roof to warn his colleagues that one of the burglars was armed. Craig fired another shot. Fairfax, still with Bentley in tow, retreated behind the only other structure on the roof, the opening from the staircase, some 30 metres from Craig.

There, Bentley surrendered his own weapons, a spiked knuckleduster and a small knife. Fairfax would later testify that Bentley told him "That's all I've got, guv'nor. I haven't got a gun."

Within minutes a large police presence and around one hundred onlookers gathered at the warehouse. Some of the police officers were armed although there are no accurate records of how many guns were taken to the scene (Yallop 1971: 61). PC Miles acquired a set of keys to the building and made his way to the roof via the internal staircase. As he emerged from the doorway, Miles was struck in the head by a bullet and fell, dying, almost at the feet of Fairfax and Bentley.

The standoff continued for several minutes. The police acknowledged firing at Craig but no accurate records were kept of which police guns were fired or of how many rounds were discharged. Bentley was taken from the roof down the same staircase that PC Miles had ascended. The police reported that he shouted "Look out Chris, they're taking me down!" (Yallop 1971: 70). Craig continued firing and yelling taunts and obscenities at the police until, his ammunition spent, he jumped from the roof. His fall was broken by a greenhouse in a neighbouring garden and he survived with a few broken bones.

The trial of Bentley and Craig began on the 9th December. Both were charged with the murder of PC Miles, both pleaded not guilty. Craig offered a *mens rea* defence – although he accepted that he had fired his pistol towards the police he had not *intended* to kill. Bentley claimed that he had not known that Craig had a gun and that he was already in police custody at the time of the fatal shot. The trial lasted just two days: ten hours of court time in total. The jury found both guilty, but made a recommendation of mercy in the case of Bentley – an educationally subnormal, epileptic young man, estimated by later psychiatric reports to

have a mental age of 11. As Craig was just 16, the judge imposed a sentence that he be "detained until her majesty's pleasure be known". Bentley was sentenced to death. Following a summary appeal, he was hanged at Wandsworth Prison on the morning of the 28th of January 1953.

Decades of controversy raged over the fact that Bentley was put to death despite having not fired a shot and having been in custody at the time PC Miles was killed. Bentley was given a posthumous pardon in 1990. His original conviction was set aside by the court of appeal in 1998. The grounds of his conviction, the rejection of his original appeal (and the disregard of the trial jury's appeal for mercy) and the reasons for the successful appeal 45 years after his death are relevant to our discussion of the role of specific beliefs and desires in establishing motives and *mens rea*.

Bentley's guilt rested on two contentions, both of which allude to what he *knew* and to his *intentions* while on the roof.

a) that Bentley knew that Craig was armed and that the younger man intended to use it to resist arrest.
b) The intention behind Bentley's alleged utterance of the phrase "let him have it Chris!"

The Lord Chief Justice of the day, the Lord Goddard, had been the original trial judge and had directed the jury that:

> … where two people are engaged on a felonious enterprise – and warehouse-breaking is a felony – and one knows that the other is carrying a weapon, and there is agreement to use such violence as may be necessary to avoid arrest, and this leads to the killing of a person or results in the killing of a person, both are guilty of murder, and it is no answer for one to say "I did not think my companion would go as far as he did".
>
> (Yallop 1971: 290)

At his trial, Bentley denied knowing that Craig had a gun. The police contradicted this. PC McDonald claimed that when he asked Fairfax about the gun, Bentley had interjected "It's a .45 Colt and he has plenty of ammunition for it". Fairfax corroborated this (with some variances) although Bentley denied ever saying it (Yallop 1971: 63). If he *had* said this,

while with Fairfax some distance from Craig, it would be fair to infer that he must have had prior knowledge; how else could he know the type of weapon and of the presence of "plenty of ammunition"?

The phrase "Let him have it Chris" has been central to the case's notoriety and was even, in part, chosen for the title of Peter Medak's 1991 feature film dramatisation of the case. Yallop (1971: 59) sums up the controversy:

> No one has been able to agree on the exact meaning of the words, "Let him have it, Chris". Did Bentley mean that Craig should give up his gun, or did he mean that Craig should offer violence? The phrase has indeed become a classic example, frequently quoted, to show the ambiguity of our language.

Lord Goddard described this alleged remark as "the most serious piece of evidence against Bentley". It implied, if interpreted as an instruction to shoot, that Bentley both knew about the gun and encouraged his accomplice to use it. Even if interpreted as an instruction to give up the weapon, it would mean that Bentley knew that Craig had it. It does not concern us here whether it was ever said – Bentley denied saying it and Craig, years after being released from jail, denied hearing it. The concern in this thesis is with how the courts interpreted the alleged remark and the inferences that were drawn about Bentley's epistemic states on the night of the shooting.

Other than alleging that Bentley had said both "Let him have it Chris" and made the remark about the gun and ammunition, the police officers were not asked in evidence to make ascriptions of specific propositional attitudes to either defendant. The evidence was restricted to what they saw, heard and said themselves. The jury was expected to infer, from the police reports of what the miscreants said and how they acted, the intentions and foreknowledge (in Bentley's case) of the existence of a firearm. In his summing up, Lord Goddard further directed the jury that:

> The great virtue of trial by jury is that jurymen can exercise the common sense of ordinary people. Can you suppose for a moment, especially when you have heard Craig say that why he carried a revolver was for the purpose of boasting and making yourself a big man, that he would not have told his pals he was out with that he had got a revolver? Is it not almost inconceivable that Craig would not have told him, and probably shown him in the revolver which he had? … I

> should think you would come to the conclusion that the first thing, almost, Craig would tell him, if they were going off on a shop-breaking expedition, was: "It's all right. I've got a revolver with me."

The judge was suggesting that the jury need not ascribe any deep insights to Bentley in order to conclude that he must have known that Craig had a gun. His suggestion is that "common sense" suggests that a boastful youth like Craig would have told him so. Lord Goddard expressly avoids the suggestion that Bentley's "common sense" would have led him to assume that Craig would be armed. He also seeks to absolve the jury of any need to interpret "Let him have it Chris" in belief-desire terms.

Following decades of campaigning, Bentley's case was referred to the Court of Appeal for a second time in 1997. The practice in hearing appeals of long-past cases is to apply the legal standards of the time of the original trial. In their findings, the appeal court judges stated that:

> In order to determine the appellant's guilt, the jury had to resolve a number of issues. They included in particular the following:
>
> 1. What was the nature and scope of the joint enterprise on which Craig and the appellant embarked?
>
> 2. When did the appellant get to know that Craig had the gun with him? None of the observations allegedly made by him were inconsistent with the knowledge having been acquired when the two were on the roof. The trial judge in the course of his summing-up to the jury suggested … that it was inconceivable that Craig would not have told the appellant when they were going on a shopbreaking expedition that he had the gun. We do not think that that is necessarily so. The appellant had no record of violence and Craig may not have wanted him to know he was armed in case he refused to accompany him.
>
> 3. Did the appellant shout out "Let him have it, Chris"? If he did, what did he intend by the words he used? In particular, it could be argued that his actions and words while on the roof thereafter were consistent with his not having wanted to incite Craig to shoot any officer and that Craig's display of hatred towards the police suggested that he was engaged on an enterprise of his own.

> 4. At the time P.C. Miles was shot, was the appellant participating or had he withdrawn from any joint enterprise that could be inferred from the evidence? Here again his actions and words on the roof were relevant and the jury would have to determine the intention behind his shout "They're taking me down, Chris".
>
> (Bingham of Cornhill et al. 1998: 9-10)

Ultimately, the supreme court Justices, led by the Lord Chief Justice of that year (and so a successor to the original trial judge) concluded that:

> For all the reasons given in this section of the judgment we think that the conviction of the appellant was unsafe. We accordingly allow the appeal and quash his conviction. It must be a matter of profound and continuing regret that this mistrial occurred and that the defects we have found were not recognised at the time.
>
> (Ibid.: 25)

The appeal was allowed explicitly not because the appeal court had concluded that the police officers who reported hearing Bentley say "Let him have it Chris" were lying nor because they interpreted those words in a particular way. The appeal was allowed because of the evidence of Bentley's behaviour, demeanour and attitude to the police on the roof that night cast *reasonable doubt* that he had any intention to resist arrest with force. This doubt would have been sufficient, the Supreme Court concluded, that *had* the trial judge correctly directed the original jury as to the burden of proof, they might have found differently.

The reductionist view of the epistemology of testimony has some bearing here. We are expected to determine whether or not we accept what we are told (the interpretation of the police officers of Bentley's intentions) on the basis of a panoply of circumstantial and background considerations (Bentley's behaviour, etcetera). The Supreme Court finding suggests that in assessing a defendant's denial of *mens rea*, we should judge whether this denial induces reasonable doubt not in isolation, nor even on the basis of inferences about their epistemic states (or their belief-desire states) from what they said, or the nature of their joint enterprise (Lord Goddard's "common sense"). We should also take into account the other salient details.

This is also true of the Morgan case. The revision of the definition of the crime of rape under the Sexual Offences Act, 2003 expressly directs that other features of the case are taken into account when assessing the value of a defendant's *mens rea* defence that they *believed* that the alleged victim had consented – including, but not limited to, whether the question had ever been asked and answered.

In establishing the *mens rea*, or evaluating a defence on that ground, the courts take a more holistic view of testimony than is suggested by Burge's "acceptance principle". Both the prosecution's allegation that *mens rea* was present, appropriate to the indictment and the defendant's denial that it was on the basis of their conflicting beliefs and desires, each stand or fall on their coherence with other evidentiary submissions. This has some bearing on how we do and how we should evaluate – our "linguistic acceptance" (Graham 2000) – of testimony about "beliefs" and "desires" when we need to establish intentions.

5.8 Chapter Summary

In two socially important areas of discourse, **justifications** and **excuse-making** (section 5.1), individuals occasionally advert to their own beliefs and desires at the time of committing the act. The kinds of act that require justifications and excuses might be simply unexpected – out of the ordinary or out of character for that individual – in contravention of some moral norm, or against the prevailing law.

Excuses are distinguished from *justifications* by the fact that an act can be *justified* only by some feature pertinent to the act and the circumstances in which it occurs. An act is justified, therefore, only when, despite initial appearances, it is in fact not anomalous, immoral or illegal. When a statement adverts to some feature of the actor's thought processes, this is an attempt to *excuse* the behaviour. Excuses accept the nature of the act but claim that some feature of the actor, of the information at their disposal or of their understanding of the circumstances (rather than the circumstances themselves) make their behaviour understandable or reasonable and so not anomalous, immoral or culpable to the same degree as it might have been had the act been fully intentional.

When an excuse is offered in belief-desire terms responsibility for the action is not usually assessed against the assumption that those beliefs and desires would have *caused* the action under examination (5.2-4). For example, no counterfactual test – "had the subject *believed* differently would they have acted differently?" – is usually applied. It is more usual that a

standard of reasonableness is applied – not only whether it was reasonable for the subject to have believed what they claim but whether their overall behaviour was in accord with the expectations of a reasonable person under the circumstances – including whether they had taken sufficient *care* given the *risks*.

The idea that care commensurate with risk plays a central role in assessing culpability is key to an understanding of the *mens rea* component of certain common law offences (5.5). Indeed, the *mens rea* component exists, according to some legal theorists, expressly so that defences can draw on the excuse that the defendant took reasonable care, despite the fact that the outcome was an illegal act (or failure to act). This point was further illustrated by means of two examples, one of which led to a change in the law (5.6) because the prior formulation permitted a defence based on *what a defendant believed*, alone. In the second (5.7) an infamous miscarriage of judgement resulted from a judge's failure to direct a jury to take into account the totality of a defendant's behaviour rather than to assume that the question of culpability rested on establishing what he knew or believed during the commission of a crime.

6 Chapter Six:
Hedges and Non-Causal Mental State Terms

*The notion of **hedges** and the linguistic function of hedging are examined by means of a number of examples. Certain uses of "believe" – especially in the first person form "I believe that …" are shown to fulfil both the criteria for being regarded as hedges and to fulfil the functions of hedges in practice. Through an examination of a second linguistic act – that of **assertion** – it is suggested that the "I believe that…" part of a hedged statement constructed this way, despite having the superficial form of an assertion, does not assert any additional content. Its role should be understood as modifying the utterer's commitment to the proposition. As such, it might be taken as referring to a mental state possessed by the utterer – perhaps that mental state could be defined similarly to a "propositional attitude". However, where this significantly differs from the view of philosophical folk-psychology is that the attitude is defined by the expressed degree of epistemic commitment and not by any specific functional or causal role in that it plays in bringing about a particular behaviour.*

The word 'belief' is a difficult thing for me. I don't believe. I must have a reason for a certain hypothesis. Either I know a thing, and then I know it – I don't need to believe it.

[asked whether he believed in God]:

Difficult to answer… I know. I don't need to believe, I know.

Carl Jung, interviewed by John Freeman on the BBC's
Face to Face television programme, 1959[98]

6.1 The Phenomenon of *Hedging and Hedges*

The contention of the second half of this thesis is that many utterances of phrases such as "I believe that …" or "I desired to do good and believed that my action was the best way to do good," serve purposes *other than the expression of an attitude to a proposition*. In this

[98] The entire programme is available online at https://www.youtube.com/watch?v=eTBs-2cloEI.

chapter I deal with another family of uses which, *although they might have the form of modifying the degree to which a speaker commits to the truth of a proposition, are principally used for other purposes*. Utterances in this class are known as **hedges** and the practice of using them as **hedging**. Hedging is the use of **modal adjuncts** or **pragmatic markers** (defined below) which make a proposition, categorization or assertion more or less *fuzzy*. "I believe that …" might be used to hedge a statement for several reasons.

Supporting these contentions requires that we look at hedges in three ways. First, I will establish three of the purposes that hedges often perform. Second, by examining how the use of hedges is described from within a neo-Gricean reconstruction of **conversational implicature** to illustrate how the addition of "I believe that" to a statement modifies the implicature in ways consistent with these purposes. Third, examining the speech act of **assertion** and its **norms** illustrates how the addition of a hedging-phrase – in particular "I believe that" – violates those norms if we assume that this phrase asserts some additional information, such as information about the specific beliefs of the utterer.

Since analytic philosophy prides itself on its commitment to precision, to entailment, and to truth, incorporating the uncertainty that characterizes so much everyday life and language-use is a challenge. Philosophy shares this challenge with linguistic pragmatics, which seeks to bring the nuances of language in the real world under some systematic description while simultaneously avoiding a proliferation of definitions.

Insights from Wittgenstein and Austin alerted some philosophers and some linguists to the expectation that language in everyday use is seldom as precise as traditional logic and grammars would demand. Inspired by Wittgenstein's remarks on family resemblances in the *Philosophical Investigations* (§67-77), theorists started to adopt the view that category membership might not be a binary question, that category boundaries might be *fuzzy*. It is always possible to make a more clear-cut case for the inclusion of some items than for others.

Zadeh (1965) both introduced the adjective "fuzzy" to describe sets to which membership was a matter of degree, and began to propose a systematic way that the resulting uncertainty could be handled. A fuzzy set was defined as a class or category of objects with a *continuum of grades of membership*. Since Zadeh, *fuzzy logic* (J. F. Baldwin 1979) and *fuzzy grammar* (Aarts et al. 2004) have become part of the landscape of investigation and of scholarship in logic and linguistics.

Directly referencing Zadeh's work, Lakoff drew attention to an established linguistic phenomenon through which language users handle fuzziness. He suggests that:

> Some of the most interesting questions are raised by the study of words whose meaning implicitly involves fuzziness - words [or phrases] whose job is to make things more or less fuzzy. I will refer to such words as 'hedges'.
>
> Lakoff (1973)

The word "hedge" can sometimes be used as a noun to refer to a precise word or form of words that perform the hedging-function or, as a verb, to an instance of performing that speech act (Schröder and Zimmer 1997). Similarly, "hedging" is sometimes used to pick out the deployment of a particular word or phrase recognised as a hedge – Lakoff (1973) included a list – or the act of making a statement more or less fuzzy (as explained in what follows). For clarity, throughout this chapter I will use "hedge" only as a verb and "hedging" to refer only to the act. When I want to refer to a particular word or form of words I will use the form "hedging-word" or "hedging-phrase". This might lead to the occasional clumsy construction but will avoid any potential for confusion.

Lakoff's list of hedging-words and phrases ranges from "sort of", "roughly" and "for the most part" to "virtually", "practically" and "in a real sense". These are modifiers that admit of *uncertainty* about the locution that they modify.

To take Lakoff's example, the category "bird" has an enormous number of members. We might be happy to admit without hesitating that a *robin* is a bird – it is a definitive member of that category. Other examples are less typical. A chicken is certainly a bird but is less likely to be thought of as the most representative example. Penguins, emus and kiwis are even further from the most readily categorised members of the set.

Interesting effects arise when a typical set member is modified with a hedging-phrase. Consider the example:

> A robin is *like* a bird

This example is less readily assented to than the unmodified statement "a robin is a bird": the hearer is likely to question the hedging phrase; in what sense is it *like* a bird? It just *is* a bird. When category membership is less clear cut, however, the addition of a hedging-word has the opposite effect;

A kiwi is *like* a bird.

Is more readily assented than "a kiwi is a bird" – especially if the hearer is unaware of the biological classification of kiwis and is simply looking at an example or an image of this rather un-bird-like creature. The addition of the hedging-word makes it clear that we are making an allowance for the fuzziness of the category. This is especially the case when the category is being extended by metaphor or simile.

A bat is a bird.

Is manifestly not true. Giving this information to someone unfamiliar with the biology of bats would be to mislead them. However:

A bat is *like* a bird.

Is *true*, at least in the sense that it is a small creature with wings and that flies. The hedging-word "like" is here acting as a **pragmatic marker**, drawing attention to the fact that the unmodified statement is *not literally true but, suitably modified, the comparison is informative*. Imagine describing a bat to a small child: "A bat is like a bird except that it has fur instead of feathers and gives birth to live young instead of laying eggs". Omitting the hedging word "like" would render this explanation as misleading, contradictory nonsense.

When Lakoff describes hedging as the act of making a statement more or less fuzzy, the effect depends on the degree to which the starting, unmodified statement is definitive. If we express the likelihood of assent to category membership as a scale from 0 – definitely *not* a member – to 1 – definitely a member – then the addition of a hedging-word or hedging-phrase renders statements that in their unmodified are closer to 1 *less* acceptable. This describes the effect of modifying "a robin is a bird" (close to 1) to "a robin is like a bird" (less likely to gain immediate assent). Objects in the middle of the range have their likelihood of attracting assent *enhanced* by the addition of a hedging-word. Thus "a bat is like a bird" is more likely to be assented to than "a bat is a bird" precisely because a bat's *characteristics* place it somewhere in the middle of that continuum from "certainly not a bird" to "certainly a bird": a bat is in fact *not* a bird, but we understand why somebody might think that it was *like* one. In the case of objects for which the likelihood of assent to category membership is close to 0, the value is unaffected by hedging. "A cow is like a bird" is just as untrue as "a cow is a bird" unless we provide additional justification for the simile.

Consider fuzzy predicates like "tall". "Simon is tall" is placed on the scale of acceptability, between 0 (Simon is definitely not tall) and 1 (Simon is certainly tall) according to facts about Simon. If he were 190cm from sole to pate we would tend to place "Simon is tall" somewhere close to 1. If on the other hand, if Simon were less than 150cm tall we would tend to place the acceptability of the phrase "Simon is tall" closer to zero. If Simon's height was measured somewhere in the mid-range, for example 175cm, then, subject to context, we would judge the acceptability of the statement somewhere in the middle of the range.

Now consider the addition of the hedging-phrase "sort of". At 150cm or less, Simon is no more "sort of tall" than he is "tall". The addition of a contextualising hedging-phrase can increase the acceptability – as in "Simon is tall *for a pygmy*" – but for the purposes of this discussion, consider the effect of the single hedging-phrase "sort of". Simon at 190cm or more is not "sort of tall"; he is just plain *tall*. The only circumstance under which "Simon is sort of tall" is acceptable would be when other pragmatic features make it plain that the speaker is being ironic or sarcastic. On the other hand, if Simon is around 175cm in height, we might quibble at the statement "Simon is tall", whereas "Simon is sort of tall" makes it clear that we are allowing for comparisons that can remain unspoken and so assent is more readily given to the hedged statement than to the unmodified form. To set out this function schematically:

> Case: Simon is 190cm in height.
> Acceptability of "Simon is tall" ≈ 1; "Simon is sort of tall" ≈ 0
> Case: Simon is 150cm in height.
> Acceptability of "Simon is tall" ≈ 0; "Simon is sort of tall" ≈ 0
> Case: Simon is 175cm in height.
> Acceptability of "Simon is tall" ≈ 0.5; "Simon is sort of tall" ≈ 1

> As any lexicographer knows, natural language concepts are fuzzy; the boundaries are not clear-cut. Zadeh has suggested that such fuzziness should be manageable, formally, in terms of what he calls fuzzy set theory. In a fuzzy set, an individual is not simply a member or a non-member, but may be a member to some degree, for example, any real number between 0 and 1.
>
> (Lakoff 1989)

This describes the *approximator function,* which is only one of the purposes to which hedges are put. I will describe this function in more detail in what follows but first we should look at the proposed linguistic role that hedging-words and hedging-phrases perform.

6.2 The Roles that Hedges Perform

Hedging-words and hedging-phrases, inserted into a sentence that otherwise asserts a proposition, can be thought of as **modal adjuncts**, which is to say that they alter the degree to which the utterer wants to commit to the truth of the proposition. For example:

> On a scale of commitment to a proposition, we might have *certainly* at the positive end and *certainly not* at the negative end, with such items as *probably*, *possibly*, *conceivably* at various points along the line, along with expressions like *perhaps*, *maybe*, *indisputably*, *without doubt*, *imaginably*, *surely*.
>
> Bloor and Bloor (2004: 55-56)

Fraser (2010) suggests that hedging should be regarded as an essential element in **pragmatic competence**. This is, he argues, an important social skill that permits a language user to "… communicate your intended message with all its nuances in any socio-cultural context and to interpret the message of your interlocutor as it was intended." He laments that the skill of hedging is seldom explicitly taught to second-language learners. A lack of facility with hedges and so with pragmatic competence more widely distinguishes second-language users from native speakers and writers. For Fraser, this restricts the ability of many language learners to convey and comprehend nuances.

Another distinction that Fraser makes is that there is "no grammatical class of hedges". Almost any word or phrase can perform one of the hedging functions, depending on the context and on the occasion in which they are deployed. We can recognise words and phrases fulfilling this functional role not on the basis of the *syntactic category* of the word or phrase but whenever we identify that it performs as a hedge. This process is a species of pragmatic inference – of which I will say more in what follows.

In a nutshell, my contention in the rest of this chapter is that the addition of the phrase "I believe that …" to an otherwise assertive utterance is an example of a *modal adjunct* as described by Bloor and Bloor (2004) in that it serves to modify the speaker's degree of commitment to the proposition. "I believe that …" also frequently marks a sentence as

communicating pragmatic inferences over and above the assertion of a proposition. It is *more than the communication of a speaker's attitude to that proposition*.

The objection to the widening of the definition of hedging and hedges from those words and phrases on Lakoff's original list is that the terms might become so broad as to become almost meaningless. Schröder and Zimmer (1997) identify hedging as a "complex research area within the fields of pragmatics, linguistics, semantics, logic and philosophy," They settle on a definition of the process of hedging as the use of language "for specific communicative purposes, such as *politeness, vagueness, mitigation* etc.," (p249, emphasis added).

From these three functions of hedges, which are paralleled in Hyland (1995a), I will examine the following three functions which hedging performs and which, I will argue, the phrase "I believe that" performs on occasions. These are:

a) The **approximator function** (vagueness). This is the hedge function that alters the likelihood that an assertion or categorisation will be assented to (in the ways outlined above). This is the function of hedges described by Lakoff (1973).
b) The **shield function**. Hedges of this kind serve to protect the utterer of from blame or censure should the asserted proposition or claim of category membership be subsequently shown to be false. This is perhaps closest to the ordinary meaning of the word "hedge" as in "to hedge one's bets". Kaltenböck (2010) argues that the shield function is the principal, although not the only, function of hedging.
c) The **politeness function**. The purpose of these hedges is to respect social conventions. The bald assertion of a proposition as fact risks appearing arrogant or failing to take account of dissenting opinions or objections. A hedge is deployed in order to avoid the possibility of giving offence. Hyland (1994, 1995b) identifies this as the principal purpose for which hedges are widely used in scientific and academic writing.

All three of these functions make use of the same effect of hedges as a modal adjunct: they all serve to alter the speaker's commitment to the truth of an asserted proposition. All hedges have this in common. This is why Crompton (1997) is moved to suggest that research into hedges should restrict itself to "language avoiding commitment, a use which corresponds closely … with the ordinary use of the word". However, this overlooks that the *purpose* that the speaker has in modifying their degree of commitment might be distinct, and that

the different purposes generate different **pragmatic inferences**. For example, by introducing vagueness to support the approximator function the speaker is implying either that "this is true to an extent" or "this is not to be taken as literally or definitively true". By deploying a hedge in the *shield function* the speaker is reducing their commitment in order to communicate "please don't blame me if this turns out to be wrong". Under the politeness function, the apparent reduction of commitment serves as a social emollient; this kind of hedge user is implying that "I am right about this, but that doesn't mean I won't respect a dissenting view or that I regard myself as superior to my audience." This explains why the use of this kind of hedge is so widespread in scientific and academic writing (Hyland 1998).

There are two dimensions, therefore, along which we can judge whether a particular word or phrase is being used as a hedge. The first is the modal adjunct dimension: identifying a word or phrase as a hedge is a matter of asking "has the inclusion of this word or phrase modified the utterer's commitment to the truth of the proposition, compared with the assertion of the proposition alone?' If it has, the word or phrase is likely to be a hedge. The second dimension concerns the pragmatic marker role of the hedge. Here the relevant question is "what additional pragmatic inferences are generated by the inclusion of this word or phrase?" This identifies the hedging function of the overall sentence.

6.3 "I believe that" as a Hedging-Phrase.

The use of "I believe that" as a hedging phrase has not widely been expressly considered in the literature. Kaltenböck (2009, 2010) analyses the use of the closely related phrase "I think" as a hedge. He identifies four functions of hedges that make use of "I think", three of which correspond to the functions identified above. "I think" is also, clearly, a modifier of the degree of commitment that an utterer admits. Much of Kaltenböck's observations about the use of "I think" apply equally to "I believe that", as will be shown in the rest of this chapter.

White (2003) draws attention to a difference in a speaker's degree of commitment between *pronouncements* – such as "I contend" or "it is a fact that" – and instances of *entertaining* a proposition – where the speaker brings in modifiers such as "perhaps", "it seems to me", "I think" and "I believe". This identifies the modal adjunct dimension of hedges and places "I believe" firmly in the camp of hedges, phrases that ameliorate the speaker's certainty. White also recognises the social (pragmatic) functions of hedges as distinct from their modal function when he writes that "…stance and attitude are fundamentally social, rather

than personal, that when speakers/writers take a stand, when they construct for themselves a particular persona or identity, it is via a process of engaging with socially determined value positions". Using a hedging-phrase, such as "I believe that …" is to take a stance relative to the hearer. This stance is distinct from that taken when the speaker wishes to reinforce their position, to make a pronouncement. It is also determined by situational factors under which the utterance takes place.

Another theorist to recognise instances of "I believe" as a hedging-phrase is Skelton (1988). He writes that "There are a very large number of ways in which one can hedge in English. Among them, for instance, are the use if impersonal phrases, the modal system, verbs like 'seem', 'look' and 'appear', sentence-introductory phrases like 'I think' and 'I believe', and the addition of *–ish* to certain (but not all) adjectives".

More systematically, Feltzer (2014) applied the techniques of corpus analysis to examine the distribution, collocates and linguistic functions of three first person cognitive verbs – "I think", "I mean" and "I believe" – in contemporary *political discourse*. The objective of the analysis was to see how often these phrases coincided with more "obvious" hedges – such as "probably", "maybe" etcetera. She concludes that:

> The local context of all three parentheticals shows a fine interplay of boosting and attenuating devices. It is characterized by the orchestrated interplay with other expressions of vagueness or fuzziness, such as pronouns with an indeterminate domain of references, adverbials or generic nouns, and with less-fuzzy making devices.

In none of Feltzer's examples does "I believe" or "I think" serve to make a claim about a particular mental state or attitude of the speaker. Their function is to mitigate or reinforce the strength to which the speaker is committing to the asserted proposition. This is a significant divergence from the role of "belief" in the understanding of philosophical folk psychology. In the traditional view, to believe a proposition is to have a particular disposition to behave in a way directed or caused by that specific belief. A declaration of a belief is to declare the possession of that disposition. This is why the ascription of belief is taken as being explanatory and predictive of subsequent action. In the case of "I believe" as a hedging-phrase, the commitment is being modified and the direction, degree and nature of the modification are determined by features of the occasion and by the pragmatic implicatures of the utterance.

To support this, I will need to suggest how such implicatures are generated.

6.4 The Pragmatic Interpretation of Hedges

Pragmatics is the branch of linguistics that is concerned with the relationship between sign systems and their users, with due regard to the circumstances of their use.[99] This entails that there may be meanings or significance to the signs that make up language that go beyond the things signified. Several researchers in the field of hedges – for example Clemen (1997), Fraser (2010), Kaltenböck (2010) and Schröder and Zimmer (1997) – characterise hedges as *pragmatic markers*, which is to say units of speech that draw attention to the fact that the meaning of the utterance may go beyond the usual meaning of the signs (words and syntax) that make it up.

The **utterance** is the unit of language that pragmatics analyses. It is an individual instance in which language is used. The same sentence of a given language might be deployed on different occasions, in different contexts by different users and with correspondingly different purposes and meanings. Each instance would be a different utterance. The objective of pragmatics is to describe the way that utterances arrive at these differences in a systematic way – without allowing an open-ended proliferation of lexical meanings.

As Recanati (1991: 99) describes it, "We have three levels of meaning: sentence meaning, what is said and what is communicated. What is communicated includes not only what is said but also the conversational implicatures of the utterance." **Conversational implicatures** is the name given to the information that is conveyed by an utterance that cannot be directly read from the meanings of its components. In Recanati's words (1991:97): "Conversational implicatures are part of what the utterance communicates, but they are not conventionally determined by the meaning of the sentence; they are pragmatically rather than semantically determined."

The central questions for pragmatics are, firstly, how are individual conversational implicatures encoded and understood (or *generated*)? And, secondly, how can we systematically describe this process in order to explain and predict what will be implicated by a particular utterance on a particular occasion – given that the variety of utterances and

[99] In contrast to *semantics*, which is concerned with the relations between signs and the entities that they signify and *syntax* which is concerned with the legitimate and systematic ways that signs relate to each other and can be combined.

occasions is potentially limitless, and assuming that some kind of system must be in play that allows speakers to control implicatures and hearers to understand them?

In an attempt to eliminate this apparent gap between language in use and the rules of logic, Grice (1991: 26) proposed that we can describe the "nature and importance of the conditions governing conversation" by reference to a number of **maxims of conversation** underpinned by a **co-operative principle**. This latter principle suggests that since language usually entails the joint efforts of speakers and hearers, language users usually endeavour to observe this rule:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.
>
> (Grice 1991: 26)

In addition to the cooperative principle, Grice proposed nine *maxims of conversation*, organised into four categories:

> Maxims of Quantity:
>
> 1) Make your contribution as informative as is required (for the current purposes of the exchange).
>
> 2) Do not make your contribution more informative than is required.
>
> Maxims of Quality:
>
> Supermaxim: Try to make your contribution one that is true.
>
> 1) Do not say what you believe to be false.
>
> 2) Do not say that for which you lack adequate evidence.
>
> Maxim of Relation (or Relevance)
>
> Make your contribution relevant and timely.
>
> Maxim of Manner
>
> Be Perspicuous
>
> Submaxims:
>
> 1) Avoid obscurity of expression.
> 2) Avoid ambiguity.
> 3) Be brief (avoid unnecessary prolixity).
> 4) Be orderly.

(Grice 1991; 26-28)

Grice recognised that there might be a need for additional maxims (Ibid: 27) if we are to provide a comprehensive account of the generation of conversational implicatures. Despite their formulation as imperatives, Grice did not intend his maxims as normative constraints on how conversations *ought* to be conducted. They are to be regarded, Levinson (1983: 101) suggests, as "a set of over-arching assumptions guiding the conduct of conversation." When a hearer notices that one of the maxims is being **violated**, or when a speaker deliberately **flouts** one or more of the maxims, each is aware that conversational implicatures are in play.

Grice suggests a further regulative dimension for the generation of conversational implicatures; the **modified Occam's razor**. Intended to ensure that the senses or definitions of a word or sentence to not proliferate beyond the bounds of usefulness, this rule is expressed as "senses are not to be multiplied without necessity," (Grice 1991: 47). Without this rule, each time that a word, phrase or sentence is used in a novel situation by a different speaker and generates a new implicature we would simultaneously generate a new ambiguity to its meaning, a new line in the dictionary. For this reason, Grice suggests, when two meanings of an utterance are possible, a literal one and a pragmatic one, we should give preference to the pragmatic, situation specific one, rather than expecting words to encode seemingly infinite literal meanings.

Armed with the co-operative principle, the maxims of conversation and the modified Occam's razor, we can systematically describe the ways in which conversational implicatures are generated by paying attention to which maxims of conversation are being accidentally *violated* or deliberately *flouted* on a particular occasion.

Grice's formulation has been immensely influential and has been developed by a number of theorists. One such "neo-Gricean" model is that suggested by Levinson (2001), as a system of **generalised conversational implicatures**. He writes:

> There must be powerful heuristics that give us preferred interpretations without too much calculation of such matters as speakers' intentions, encyclopaedic

> knowledge of the domain being talked about, or calculations of others' mental processes[100].
>
> (Levinson 2001: 4)

This is the contrast between Levinson and Grice. Although Grice's *maxims of conversation* form the foundation on which Levinson's **heuristics of implicature** are built, Grice regarded inference to a speaker's intentions – including their mental processes – as essential to communicative success (Grice 1991: 86-116). Levinson holds that his implicatures are generalised and presumed, such that the series of heuristics that he proposes unlock default interpretations. Thus they generate *generalised conversational implicatures* independently of any inference of a speaker's intentions.

Levinson proposes three principles:

1. If the utterance is constructed using simple, brief, unmarked forms, this signals business as usual, that the described situation has all the expected, stereotypical properties;
2. If, in contrast, the utterance is constructed using marked, prolix or unusual forms, this signals that the described situation is itself unusual or unexpected or has special properties;
3. Where an utterance contains an expression drawn from a set of contrasting expressions, assume that the chosen expressions describe a world that itself contrasts with those rival worlds that would have been described by the contrasting expressions.

(Levinson 2001: 6)

From these principles, Levinson develops a typology of general conversational implicature based on three *heuristics of pragmatic inference*.

Under the **Q-heuristic** (from *quantity*) "what isn't said isn't the case" (Levinson 2001: 35). Levinson notes that this heuristic is straightforwardly derived from Grice's first *maxim of quantity*: "make your contribution as informative as is required (for the current purposes of

[100] Transpose this observation from the field of linguistics to the area of action explanation and philosophy of mind and psychology and it could be a manifesto for this thesis!

the exchange)". This heuristic, in common with the maxim, generates scalar implicatures. For example:

> Michael and Mary have three children.
> +> (conversationally implicates) Exactly three children, not four, five or more.

Levinson's second heuristic is the **I-heuristic** (informativeness): "what is expressed simply is stereotypically exemplified". This relates to Grice's second *maxim of quantity*: "Do not make your contribution more informative than is required. Levinson argues that this heuristic allows us to differentiate between:

> fruit bat/cricket bat/aluminium bat.

Although these three locutions are syntactically similar they generate distinct implicatures in simple usage thanks to their stereotypical associations. If we wanted "fruit bat" to designate either "a piece of sporting equipment made of fruit" (cf. "aluminium bat") then more complex utterances would be required to generate the non-stereotypical interpretation. Levinson (2001: 37) also suggests that this allows the derivation of a bi-conditional implicature from a simple conditional:

> If you mow the lawn, I'll give you five pounds.
> +> If, and only if, you mow the lawn I will give you five pounds.

The third heuristic suggested by Levinson is the **M-Heuristic** (from *manner*); "what's said in an abnormal way isn't normal". This derives from Grice's *maxim of manner*: "be perspicuous" – and particularly from from the first ("avoid obscurity of expression") and third ("be brief; avoid unnecessary prolixity") *submaxims*. It is the converse of the *I-Heuristic*; it suggests that non-simple usages – for example, those laden with additional words or unusual formulations – are to be interpreted non-stereotypically. Levinson (2001: 39) offers the example of the deliberate use of a double negative, such as:

> It's not impossible that the plane will be late.

Which implicates that the plane is less likely to be late than would be suggested by:

> It's possible that the plane will be late.

This is an illustration of how conversational implicature allows a speaker to make fine distinctions and for those distinctions to be understood. Similarly, if we describe someone as being:

> Not entirely uninterested in football.

The implicature is that this person is something of a football obsessive. Another example of a prolix or obscure usage generating a non-typical implicature, offered by Levinson (2001: 37):

> The corners of Sue's lips turned slightly upward.
> +> Sue didn't exactly smile.

To reiterate Levinson's three heuristics:

> *Q-Heuristic*: What isn't said, isn't [the case].
>
> *I-Heuristic*: What's expressed simply is stereotypically exemplified.
>
> *M-Heuristic*: What's said in an abnormal way isn't normal.

As with any heuristic, the application of these generates non-entailed inferences. Errors of commission and interpretation are still possible, even when the heuristic rules are followed. Levinson argues only that the implicatures generated by these heuristics have "the status of preferred interpretations, because the heuristics will be understood to be generally in force – it is that which gives them their communicational efficiency." (Levinson 2001: 39).

Complex implicatures emerge from interactions between Levinson's heuristics. To deal with this, Levinson proposes a hierarchy or "ordered set of priorities" through which inconsistencies can be resolved. Where there is a conflict between potential inferences we should, he argues, assume that inferences from the *Q-heuristic* defeat those derived from the *M-heuristic* which in turn defeat those generated by the *I-heuristic* (Levinson 2001: 39)

6.5 Implicatures of "I believe that…" Hedges

Statements that begin with "I believe that" or that include a parenthetical "I believe" are hedges if they modify the speaker's degree of commitment to the propositions they contain. Determination of their function – differentiation between, for example, *approximator*,

shield and *politeness* functions – requires the generation of further implicatures as can be seen if we apply Levinson's heuristics to such utterances.

Under the *approximator function*, the purpose of the hedging-phrase is to render the hedged statement more vague and thus more acceptable if it is in the mid-range of certainty and less acceptable if it is at the upper end of the scale. For example, the proposition:

> There are more than seven billion people on earth.

Stating this proposition (a well-attested fact) suggests, by application of Levinson's Q-heuristic suggests the following scalar implicature:

> +> There are more than seven billion people on earth and this is the closest round number.

"More than seven billion" includes, literally, eight, nine or twenty billion. However, by the lights of the Q-Heuristic it would be perverse to claim of a world with a population of twenty billion, that there are "more than seven billion" – unless there was an express reason for using that number – even though the statement would be true. Compare the statement:

> *I believe that* there are more than seven billion people on earth.

Immediately the speaker is indicating some uncertainty about the number, signalling that it is a *hedge*. This is clear because, *pace* the traditional *propositional attitude* reading of a statement beginning "I believe that", by Levinson's *I-Heuristic* (what's expressed simply is stereotypically exemplified) the statement "There are more than seven billion people on earth" conversationally implicates "I believe that there are more than seven billion people on earth". When the speaker adds the unnecessary additional "I believe that" they are, by the application of the *M–Heuristic* ("What's said in an abnormal way isn't normal") signalling something else. What, precisely, will depend on the context and on the function that the hedge is deployed to perform? The speaker may simply be recording that the number is imprecise – although the implicature generated by the *Q-Heuristic* is unaffected by the addition of "I believe that". The number being spoken of is *still* seven billion and not twenty billion. The speaker might be seeking to shield themselves from the possibility that this dynamic number has passed eight billion since the last time they heard of it and so avoided censure or loss of face associated with a genuine mistake. Alternatively, in a formal group the speaker will avoid appearing brash, overconfident or rude by asserting this

knowledge if they ameliorate it with "I believe that" – even if they are quite certain of the knowledge.

The function will affect the full implicature and the function is given by the context and to occasion.

To illustrate this, take a more developed scenario. A caterer has been commissioned to provide canapés for a reception. Their delicious prawn vol-au-vents are disappearing more rapidly than expected. The nervous host asks the caterer whether there will be enough for all of the guests. The caterer might answer:

> I made enough for the expected number of guests.

Which is a simple assertion and, again, by the *I-Heuristic*, implicates that the caterer *believes* what is being asserted. In the event the answer given is a hedged one:

> *I believe that* I made enough for the expected number of guests.

We could understand this hedged statement as a straightforward statement of uncertainty about the number (approximator function) although it is unlikely that a responsible caterer would not know how many of a particular item they had provided. It could also be read as an instance of *politeness:* the host is the caterer's employer, after all. Much more likely, however, is that this hedge is being offered as a *shield*. The implicature is something like:

> +>I know that I have provided enough of this item for all of your guests to have some. It is not my fault if some of them have taken more than their fair share.

This implicature is generated both by the situation and by Levinson's *M-Heuristic* (what's said in an abnormal way isn't normal). Adding those three words indicates not only that the caterer is defensive about having done their job professionally but also that any shortages are the fault of the host's voracious guests.

For a third example, consider a situation in which an esteemed professor of mathematics is expounding on a detailed proof in front of a class of junior undergraduates. He marks up an example on the blackboard but, unwittingly, makes an obvious error in the demonstration. One or two students raise a nervous hand, the professor barks at one of them, demanding an explanation for this unwelcome interruption. Two ways are available for the student to express their concern:

> The professor has made a mistake there.
>
> *I believe that* the professor has made a mistake there.

Again, by the *I-Heuristic*, the first statement implicates the second, so the second form is a non-stereotypical formulation and, by the *M-Heuristic* must carry an additional implicature. I contend that in this scenario, the second formulation is much more likely to be the student's choice of words. This is not because they need to express uncertainty: the professor's error is a glaring one, plain for the even freshest undergraduate to spot.

This is a clear example of the *politeness function* of hedging. In saying "I believe that the professor has made a mistake" rather than merely stating the proposition, the student is respecting the authority of the professor and the power distance between them. Had a student made a mistake, then the professor would be much more likely to assert the non-hedged form "You have made a mistake there" – although the polite, hedged form would be equally acceptable and the implicature of the hedge would still be clear as an instance of politeness – of not wanting to undermine the student in front of their classmates.

This kind of social function is also an example of the *I-Heuristic* at work. In the absence of the stereotypical interpretations of the hedge as either an approximation or a shield, we read the hedge as a social emollient. It is, however, by the *I-Heuristic*, a stereotypical way for a student to address a professor and generates the following implicature:

> I believe that the professor has made a mistake there.
> +> I am respectfully pointing out an error without, in any way, seeking to challenge the professor's authority.

The uses of hedging in this social or politeness function are not restricted to occasions of disparity of authority or power. Such usages are almost ubiquitous on academic writing – including in the sciences where one might expect fidelity and precision to be at a premium. Hedges occur so frequently in this domain that Myers (1989) is prompted to write that "the hedging of claims is so common that the sentence that looks like a claim but contains no hedging is probably not a statement of new knowledge".

Consider these two sentences:

> We believe that this shows that the addition of reagent x increases the speed of the reaction.

> This shows that the addition of reagent x increases the speed of the reaction.

The first is the hedged statement, but the hedge does not express the author's lack of commitment to the conclusion. One expects that any scientific paper published in a reputable peer-reviewed journal represents the best available data and the most secure inferences available to its authors. The implicature of the hedged statement is thus:

> +> We respectfully commend to our peers and to the wider scientific community this novel conclusion – that the addition of reagent x increases the speed of the reaction.

This implicature is, once again, generated by the *I-Heuristic*; politeness in scientific and other academic research articles is the usual mode of expression (Crompton 1997; Hyland 1995b, 1995a, 1996, 1998; Myers 1989; Salager-Meyer 1994) so that it becomes the stereotypical mode of expression. So important is the politeness function that Leech (2003) has proposed adding a "principal of politeness" to the Gricean maxims. Huang (2007: 37n) rejects this on the grounds that it invites an unnecessary proliferation of maxims and that politeness is a social, cultural phenomenon rather than a linguistic one. Levinson has gone on record as acknowledging the importance of politeness phenomena to pragmatic understanding but suggests that this can be dealt with under his heuristic schema (Levinson 2001).

6.6 "I believe that…" Hedges and Assertion

All of the examples in the previous section shared one feature; the hedges modified an **assertive statement**. Indeed, the hedged form could also be taken as a form of **assertion**, even though the content of the assertion was modified by the hedge and by the resultant conversational implicatures. Another way to examine the phenomenon and its particular implications for uses of the phrase "I believe that" is by reference to contemporary work on the linguistic phenomenon of assertion.

Superficially, a bare assertion of a proposition, p, and a hedged utterance of the form "I believe that p" are both assertions. We have seen that, by the application of Levinson's *I-Heuristic*, the bare assertion will generate the implicature that the utterer believes the content of the proposition. Given this, we should examine whether an utterance where "I believe that" is made explicit asserts anything different or additional to a statement of the bare proposition. Either it asserts additional information – for example that the speaker's attitude to the proposition is one of belief (which, according to Levinson's *I-Heuristic* would

be redundant), or it changes the conversational implicature of the overall, hedged utterance – how we should understand what is being asserted. This latter possibility raises the question of whether hedged utterances can ever qualify as assertions. To answer this, we will need to look at the way assertion is currently understood and the **norms** that some theorists have proposed to govern this speech act.

In the introduction to a recent collection of papers on philosophical approaches to assertion Jessica Brown and Herman Cappelen suggest that "the notion of assertion has played an important role in the philosophy of language for the last 100 years". Despite this heritage, they note, providing a definitive account of the features that differentiate assertions from other speech acts remains a matter of heated debate. They list five possibilities:

> (i) Assertions are governed by certain norms – the norms of assertion.
> (ii) Assertions are those sayings that have certain effects.
> (iii) Assertions are those sayings that have certain causes.
> (iv) Assertions are those sayings that are accompanied by certain commitments.
> (v) There is no one set of sayings (of declaratives) that is correctly characterized as the set of assertions. Sayings are governed by variable norms, come with variable commitments and have variable causes and effects. There can be no substantive debate about which of these subsets are the assertions.
>
> (Brown and Cappelen 2014: 3)

As an illustration of what is at stake in these distinctions, consider the case of *telling a lie*. In ordinary terms this might be described as "asserting a proposition while knowing it to be untrue". According to norm-based accounts (i), lying is a violation of the norms and so should be disqualified from being regarded as an instance of assertion at all – lying, on that reading, is an altogether different speech act from asserting. According to the effect-based account, however, if the effect of an assertion is taken to be to cause a hearer to accept the asserted proposition as true, then lying is a species of assertion after all.[101]

As Brown and Cappelen point out, the fifth class of assertion theories (v) "rejects the assumption that there is a unique, correct way of picking out assertions from sayings"

[101] This is not to say that all norm-based accounts would disqualify lying as a kind of assertion nor that all effect-based accounts would include it. My intention is only to illustrate the significant differences that can arise from these different accounts.

(Ibid.). "Assertion is largely a philosopher's term, and we can, for different purposes, use it to pick out different subsets of sayings." (Ibid.: 4)

However, they also insist that "one common core in theories of both implicature and presupposition is that they *contrast* with assertion. What is asserted is not presupposed and it is not implicated." This would entail that any utterance that contains both asserted propositional content and content that derives from an implicature cannot be regarded as *asserting* both sets of content. Only the matter that qualifies as an assertion (under whatever characterisation one chooses) is asserted content. The content of the implicature is distinct from this. In cases where "I believe that" serves as a hedging phrase and so gives rise to hedge-based implicatures – both to modify the commitment of the speaker to the proposition and other implicatures dependent on the function of the hedge (approximation/shield/politeness) – the contents of the implicatures are not asserted.

For the sake of simplicity – and because the argument that I want to make does not directly depend on which kind of distinction one wants to make – I will consider the impact of hedging in the light of theories of the norms of assertion (i).

According to Williamson (1996) and others, the speech act of assertion is governed by certain norms. Even before we attend to the business of defining what the specific norms of assertion are, there is controversy over their character. Norms of assertion might be **alethic**, which is to say dependent on the truth of the proposition being asserted, or they might be **epistemic** – dependent on the knowledge of the person who makes the assertion (Maitra 2014: 277). Norms can also be distinguished as to whether they are *regulative* or *constitutive*. Violation of a regulative norm would imply that the speaker asserts defectively; violation of a constitutive norm entails that they fail to assert at all (ibid.).

Candidates for the norms of assertion include:

> *Truth Rule*: One must assert p only if p is true
>
> *Warrant Rule*: One must assert p only if one has warrant to assert p
>
> *Knowledge Rule*: One must assert p only if one knows p
>
> *Belief rule*: One must assert p only if one believes that p
>
> (Brown and Cappelen 2014)

Of these, the Truth and Warrant rules are *alethic norms* while the Knowledge and Belief rules are *epistemic norms* (although "knowledge" entails an alethic component). As written, they are ambiguous as to whether they are regulative or constitutive although this status could easily be made more explicit: for example, the truth rule could be reformulated as "one asserts P only if P is true" (constitutive) or the belief rule might be rendered "one *should* assert P only if one believes that P". Brown and Cappelen (2014: 3) admit that "Those who endorse such views vary in how they think about the nature of the norms and what it means to say that we follow these rules."

There are other candidates. For example, *purposive norms*, which are norms that relate to the idea that assertions are distinguished by norm and by effect, a hybrid of theories in categories (i) and (ii). A speaker asserts (or asserts well) *only* if they speak with a suitable assertive purpose – such as persuading a hearer of the truth of a proposition. Maitra (2014: 282) endorses this view, claiming that "neither the knowledge nor the truth norm tells us what it is to assert something. Rather, they each assume that there is something that counts as asserting, and tell us at what an asserter ought to be aiming when performing the speech act." She also insists that the content of an assertion must be propositional and so truth conditional because "a speaker who does not aim to say anything truth-conditional does not count as asserting at all," (Ibid: 292). This would entail that the truth-conditional nature of the asserted proposition (as distinct from its truth) is a constitutive norm of assertion. This is, of all the candidates for norms of assertion, relatively uncontroversial. Assertion is, if nothing else, about communicating information that is either true or false.

For the purposes of this discussion I will assume these minimal features of assertion

a) Assertion is a speech act governed by certain norms.
b) Any utterance that violates those norms either fails to assert altogether (in the case of constitutive norms) or asserts defectively (in the case of regulative norms).
c) The minimal constitutive norm of assertion is that any assertion must communicate truth conditional (propositional) information.

Accepting c) as a minimal constitutive norm does not commit one to norms of truth, warrant, knowledge or belief nor to any constitutive norm. It can be admitted that c) might be necessary but not sufficient for an act of assertion – there are other speech acts that convey truth-conditional information, such as *promising, commanding* or even *justifying*. If the

information contained in the statement is *not* either true or false, however, one is not asserting but doing something else.

In a statement of the form "I believe that p", the content of "p" is a truth-conditional statement. The questions at issue are *whether saying "I believe that p" is the same speech act as asserting p, and whether it is still an assertion – even if in being spoken it asserts both p and something else*.

Take the case of a group of friends out for a hike in unfamiliar country. They come to a fork in the path. After several minutes of debate and much fumbling with maps one of them asserts that:

> The path to the left will take us where we are heading.

Note that under the *knowledge norm* of assertion this qualifies as an assertion *only* if the speaker *knows* this to be the case (if constitutive) or is *asserting defectively* (if regulative) *if they do not know*. By the *belief norm* the test of asserting or asserting well is only that the speaker *believes* what they say (that is, it doesn't have to be *true*). Under both of these norms the speaker is taken to be holding the content of the assertion to be true, the difference is in whether it has to be true. This is also true of the truth-norm, and is guaranteed by Levinson's *I-Heuristic* since we take all assertions to be a statement of whatever the speaker holds as true. Grice's *Maxims of Quality* also warrant the notion that a speaker will usually be trying to make their contribution one that is true, not say something that they know to be false nor offer anything for which they lack evidence (c.f. the *warrant norm*). This latter constraint suggests that *if challenged they would be able to offer some justification for the view asserted*:

> The path to the left will take us where we are heading.
> *Why do you say that?*
> I can see from the map that it's the shortest route.

However constituted, the assertion generates the implicature that:

> +> *I hold it to be true that* the path to the left will take us where we are heading.

The implicature is not an additional assertion. It is a component of the speech act of assertion. On the other hand, the speaker might choose to *hedge* their statement with the addition of:

> *I believe that* the path to the left will take us where we are heading.

Superficially, this addition of three words makes the implicature of the bare assertion explicit. However, taken as a hedged-statement, the possible overall implicatures are quite different:

> I believe that the path to the left will take us where we are heading.
> +> I am not totally sure (I don't know) but we should probably take the left path.
> (hedging as modification of commitment)
> +> Don't blame or condemn me if I am wrong but we should take the left path.
> (hedging under the shield function)
> +> I know that everyone else here can read a map as well or better than me and I don't want to impose my will, but in the interests of moving on we should take the left path.
> (hedging under the politeness function)

Which hedging function is being performed will be clear to the hearers on the basis of other cues, including the speaker's tone of voice and aspects of their personality traits (different for a dictatorial type from someone who is more of a team-player, for example).

This raises an issue as to whether the hedged statement, in order to function as a hedge, needs to be taken as *asserting a further truth about the mental states of the speaker*. To do so would imply the following implicature:

> I believe that the path to the left will take us where we are heading.
> +> *I hold it to be true that I hold it to be true* that the path to the left will take us where we are heading.

The speaker might, conceivably, be drawing attention to the fact that they are in possession of this specific mental state: however, in doing so, they would be drawing attention to the fact that they have a reduced level of certainty or, in the case of the politeness function of hedging, that they are giving the *impression* of a reduced level of epistemic commitment in order to avoid seeming arrogant.

At this point the proponent of the belief-desire view of action, of the belief-desire law or philosophical folk psychology more generally might be tempted to take issue. Surely, they might argue, if "I believe that" serves to indicate (or even to mimic) a reduced degree of

certainty compared either to a bare assertion or to a statement beginning "I know that", does this not entail that "*believe*" refers to a particular *epistemic mental state*? Does this not vindicate the view that "belief" and, by extension, "desire" refer to *precisely those functional mental states demanded by philosophical folk psychology*?

I have never denied that "belief" might, on occasion, refer to a particular epistemic attitude an equivalent of holding a proposition to be true. Neither does my *modest eliminativism* rest on the contention that such states do not exist. However, in all of the hedging-functions described in this chapter, the use of "I believe that" as a hedging phrase tells us *nothing new*, about the **causes**, real, supposed or potential, of the speaker's actions. Our understanding of the *causes* of a person's actions always rest with aspects of the situation, their personal traits, the meanings that they associate with their understanding of the situation and the emotional impact of these meanings. None of these things is impenetrable which is why we are usually successful at interpersonal understanding.

The addition of "I believe that" as a hedging-phrase to the statement (or assertion) of a proposition provides additional information in the form of modifying their genuine or simulated commitment to a proposition. It reduces their liability to be blamed or establishes their sensitivity to the social situation. Hedging with "I believe that p" adds nothing to our expectation or explanation of their future or past actions compared to the same individual, in the same situation, *asserting* that p. *It has not picked out a cause*.

6.7 Chapter Summary

Hedging describes the strategy in language use of deploying additional words or phrases to make a concept or statement more or less *fuzzy* (Section 6.1). Typical hedging-words or hedging-phrases include "probably", "sort of"," like", "in a way", "to some extent" or "actually". Although all hedges have in common that they either modify or appear to modify a speaker's commitment to a proposition, claim or assertion, in practice, hedges can fulfil a number of functions, among them *approximation*, *shielding* and making an utterance more *polite* (6.2).

Certain *cognitive verbs* can serve as hedging-phrases. Among these are "I think", "I mean" and, importantly "I believe" (6.3). The modification of an otherwise *constative* phrase by the addition of a hedging-word or hedging-phrase alters the *conversational implicature* of the overall phrase compared with the bare constative or *assertion* (6.4). The precise content

of the implicature depends, to some extent, on the function that the hedge is performing: the generated implicature will be different, for example, if the hedge is used to shield the speaker, compared to the implicature of a hedge deployed to fulfil the *politeness function* (6.5). Nevertheless, the same rules for the detection of conversational implicature apply to hedges in all their functions. This is true whether the rule set is based on Grice's Maxims of Conversation or the Neo-Gricean schema of heuristics proposed by Levinson.

Using the latter schema makes it clear that reference to the beliefs of the speaker is conversationally implicated in any constative or assertion of a proposition. This is why the addition of an initial or parenthetic "I believe that" or "I believe" can be understood as a non-typical formulation and interpreted as a hedge.

Understanding that hedge-phrases that apparently refer to the beliefs of the speaker do not make a claim further to the proposition being modified is underlined by an appreciation of the *speech act of assertion* (6.6-7). In particular, if assertion is governed by *norms* (6.7), then the addition of an explicit statement of what the speaker believes is disqualified from being an assertion further to the bare proposition. Whether or not hedged sentences are altogether disqualified from being satisfactory assertions or from being assertions at all will depend on the precise set of norms of assertion with which one chooses to work. In any case, and to reinforce the point made by discussion of conversational implicatures, altering the speaker's commitment to the truth of an asserted proposition is a distinct speech act from the making of an assertion. Regarding hedged phrases beginning "I believe that" as asserting something already implicated by the bare proposition leads to the absurdity of repeated reference to what the speaker holds as true.

Despite the fact that using "cognitive verbs" as hedging phrases suggests that they refer to specific mental states, this does not entail that they refer to the *causes* of an action – as is demanded by philosophical folk psychology. Thus the *elimination of "belief" from causal accounts of action* remains a live possibility, even if beginning a sentence with "I believe that" is understood as modifying the speaker's epistemic commitment to what follows.

7 Chapter Seven:
Conclusions and Implications

Abstract: In this concluding chapter, I review the objectives and the material presented throughout this thesis, present the arguments suggested by this material and suggest some future directions of research arising.

Know then thyself, presume not god to scan;
The proper study of mankind is man.

Alexander Pope, 18th Century.

7.1 Objectives Reviewed

Why do people choose to act as they do? How do most people, most of the time find the actions of other people – and their own actions – explicable and predictable? And when we give a reason for an action, what relation does the reason bear to the causes of our action choice?

This thesis seeks to address these and other questions related to human action-choice, interpersonal understanding and reason-giving. It takes inspiration from Gordon Baker's understanding of Wittgenstein's injunction in the Philosophical Investigations (PI §116) that we seek to clarify philosophical problems by bringing "words back from their metaphysical to their everyday use" (Baker 2004). Specifically, my contention is that a clarification of the implicit commitments in the way that "belief" and "desire" are used in philosophical folk psychology can help to avoid various philosophical problems that arise from that picture. This clarification can be facilitated by:

a) Paying attention how the phenomena that are the subjects of philosophical folk psychology are dealt with in cognitive and social psychology
b) taking clearer and more inclusive view – what Wittgenstein at PI §I22 describes as a "perspicuous representation" (Baker 2004: 22) – of the everyday uses of "belief" and "desire" and the application of related concepts.

This goal yields the two guiding questions of the thesis:

GC1: How does scientific psychology account for action choice and interpersonal understanding and do these accounts employ "belief" and "desire" in the causal functional roles suggested by philosophical folk psychology?

GC2: How are "belief" and "desire" used in everyday discourse other than referring to specific causal-functional *mental states* as suggested by philosophical folk psychology?

A motivator of the approach is the sense that some philosophical problems, reviewed in the introduction, arise directly from the *picture* entailed by philosophical folk psychology. A **philosophical picture**, in the sense that I use it, is an inflexible, fixed view of a phenomenon which persuades the philosopher in its sway to *commit to a number of unwarranted, perhaps tacit and unacknowledged assumptions* which arise from the metaphysical appropriation of everyday terms like "belief" and "desire". A philosopher can be led to question the picture by being shown how *some uses of the terms and concepts from which the picture is constructed do not necessitate the same metaphysical uses*. After Morris (2007), I take it that the acceptance of a philosophical picture risks tying philosophers to a **dogma** or **prejudice**. Dogmas and prejudices are unwarranted because they have not engaged with scientific or everyday uses of the terms and concepts. After Williams (2001), where a particular "default" view (in this case, belief-desire psychology) is held without warrant, it can be subject to challenge by the presentation of contrary evidence. Although holding the default view without warrant might be justified, additional warrant is required if the challenge is to be resisted.

The objective is not to refute the positions inherent in the picture, but to loosen their grip by challenging their certainties.

In the following review I outline how I have sought to challenge these commitments by presenting alternative pictures of the target phenomena drawn from cognitive and social psychology and from specific applications of "belief" and "desire" in everyday discourse. I also set out the key arguments of each chapter and how these relate to these assumptions.

7.2 Review

Part One of this thesis takes examples from two disciplines within contemporary scientific psychology to see how they address the three phenomena (actions, interpersonal understanding and reason giving) that the belief-desire law sets out to explain.

Chapter One sets out to describe the causal-explanatory goals and methods of psychology. Of particular interest were the observations that psychology seeks **genuine explanations** that describe nodes in causal networks by means of **functional analysis**. There is little room for law-like regularities and if these did feature it would be as the **explananda** of the discipline. Two examples of functional analysis are offered: firstly, the investigation of the human visual system in **Marr (2010)**, a paradigmatic example of functional analysis from the psychological through computational modelling to the neurological level. Secondly the historical development of the concept of **spreading activation in semantic memory** demonstrates how personal-level psychological phenomena can be analysed into sub-personal functions in the development and refinement of an explanatory model.

The principal arguments presented in Chapter One are:

1) Psychology seeks *genuine causal explanations* of phenomena rather than "as if" speculations based on regularities.
2) Psychological explanations do not rely on the formulation of laws. Law-like regularities are the *explananda* of psychological investigations.
3) Genuine explanations entail **functional analysis** whereby complex functions are described in terms of the more simple (sub-personal) functions that produce them.
4) Simple functions related in systematic ways bring about complex functions – and so are the causes of those complex functions.

These arguments yield the two conditionals which motivate **Chapter Two** of the thesis, and a third that motivates **Chapter Three**:

> If the possession of specific beliefs and desires are the *causes* of specific actions (metaphysical commitment 1) then we would expect them to feature in causal explanations offered by **cognitive psychology** (in the case of judgement and decision making leading to action choice).

This would mean that at some level of functional analysis, causal beliefs and causal desires would feature.

Also:

> If the **belief-desire law** holds, then we would expect the relationship between the possession of specific beliefs and desires and consequent actions to be just the sort

of law-like regularity, or explanandum, that psychologists would seek to functionally analyse.

In **Chapter Two**, the *idealised picture of reasoning* outlined by the belief-desire law or by its close relative **expected utility decision theory** seems to describe neither real-life decision making nor the prevalent understanding of judgement and decision making in contemporary cognitive psychology. Many of our judgements and decisions appear to be based on the deployment of a number of adaptive tools, including **heuristics** which are *rules of thumb that deliver reasonably accurate judgements in the contexts in which they are made.* This contrasts with normative rules that determine or constrain what is right or reasonable. The occasional errors and biases where heuristic judgements conflict with normative rules are not failures of our reason, they are a feature of the way that our decisions are ecologically suited to real-world conditions (Gigerenzer and Brighton 2011).

Cognitive psychologists also now believe that a great many of our judgements – perhaps the overwhelming majority – are arrived at **automatically**, outside of conscious awareness or control. When we ask an individual to describe the thought processes – the specific beliefs and desires, for example – that led them to a particular choice we are *inviting them to rationalise ex post facto*. Work on **automaticity** and **heuristic reasoning** have shown that, in contrast to the view that beliefs and desires are, in principle, consciously accessible, many actions, judgements and decisions take place without the actor being aware of the influences that bring them about.

The arguments developed in Chapter Two:

1) Belief-desire psychology and Expected Utility Decision theory amount to much the same thing.
2) Both of these might be normative prescriptions of idealised rationality but they are not descriptive of what actually happens.
3) Competing accounts of the way that people reach decisions, including the use of heuristics and automatic judgements are more descriptively accurate.
4) Given that people live in the real world and that their judgements and decisions have real-world consequences, perhaps a normative model based on **ecological rationality** would lead to better decision strategies.

The conditional of **interpersonal understanding** that motivates Chapter Three can be rendered as:

> If the prediction and explanation of action rested principally on the ascription of specific beliefs and desires to an actor, as suggested by the belief-desire law, then **attribution theory** in social psychology, which is concerned with how individuals (the 'folk' of folk-psychology) understand and describe the causes of action, would principally be concerned with investigating the capacity to infer and to ascribe beliefs and desires.

A central contention of philosophical "folk psychology" is that it *is the way that ordinary people explain and predict each other's actions*. Contemporary social psychology covers much the same territory under the heading of **attribution theory**. This field is directly concerned with the strategies that people use to attribute causes to behaviour in order to understand why people act as they do.

Attribution theorists investigate two kinds of attribution (Heider 1958). **Situational attributions** assume that the *cause* of an individual's behaviour or choice of action is located in the *features of the situation under which they act*. **Personal attributions** locate the cause of a particular behaviour in features of the agent. If the agent is known to the person who makes the attribution, they will advert to "what the person is like" or other **stable personal traits** in preference to all other causal explanations. However, this is also where significant biases – such as **stereotype activation** ("people like that always act that way") can override other causal attributions. *People exhibit a bias favouring personal over situational attributions even when the situational constraints over personal choice are made explicit*. This is known as the **fundamental attribution error** or **correspondence bias**.

Heuristics and automatic judgements have also been shown to play a significant part in interpersonal understanding. Heuristic reasoning has been suggested to account, in part, for another bias in interpersonal understanding – **the actor-observer difference**. Put simply, this bias suggests that people are more likely to attribute their own actions to situational causes and more likely to attribute other people's actions to personal or dispositional causes. Part of the reason for this might be what is salient to the person depending on whether they are an actor or an observer. The actor is grappling with the question "what kind of situation is this?" while the observer is able to ask "what kind of person is that?". This is significant because under the belief-desire model people are thought to understand their own and other

people's actions in the same way – by ascribing specific propositional attitudes. This is also challenged by work in social psychology on the **unreliability of introspection**.

The principal arguments of Chapter Three are:

1) People are prone to attribute agency even in situations where no agent is present (Heider and Simmel 1944).
2) Attribution theory suggests that when attributing causes to agents, people default to features of the situation or of the agent rather than considering an agent's beliefs and desires.
3) Certain systematic biases in attribution strategies indicate that individuals do not adhere to normative rules when explaining or predicting actions, but use heuristic rules including recognition, similarity and fluency. Specifically:
    a. The *Fundamental Attribution Error/Correspondence Bias* suggests that people are prone to seek features of the actor that explain their behaviour even when situational factors are both sufficient and most salient.
    b. *Actor-observer asymmetry* suggests that people are prone to attribute situational factors as the causes of their own action whereas the actions of others are usually put down to stable traits of that person.

Overall, **Part One** of this thesis lays out the contention that although cognitive and social psychology are committed to genuine, causal explanations, the models most in use in contemporary approaches to both judgement and decision making (and so action choice) and causal attributions (and so interpersonal understanding) do so largely without recourse to belief-desire ascriptions. These alternative ways to understand action and action explanation *do not share the commitments of the philosophical folk psychology picture*.

Part Two of the thesis discusses some surprising ways that the terms "belief" and "desire" feature in everyday use without referring to the presumed causes of action.

Chapter Four opens Part Two with a discussion of the ways that actions are described and explained within **narratives**. I suggest that the *construction of narratives is critically important to our understanding of time, events and action*. Examining **narrative psychology** and **narrative therapy** highlights the fact that both make use of narratives – including narratives in which beliefs and desires might feature – *without ascribing causes to specific beliefs and desires*. Fictional narratives in which the specific beliefs and desires

of a character are made explicit are weaker thanks to the fact that they are unlike real life. We do not encounter people in our everyday lives attached to cloud-like bubbles from which we can read their thoughts. We expect our fictional narratives to relay information in much the same way – revealing the character through their action choices rather than vice-versa.

We are also sensitive to a contrast between **narrative truth** and **causal, historical truth**. The former plays a crucial role in the way that we construct meaning around events and occurrences. The latter is essential to a causal account such as that demanded by the scientific approach.

That the kind of narrative proposed by the belief-desire picture might be culturally dependent is supported, firstly, by an exposition of action-description in the Homeric epics. These foundational texts of western European literature *make no use of propositional attitude ascriptions* but describe people's actions in terms of *their hearing voices or having bodily sensations that compel them to act*. Another largely non-literate contemporary culture, the *Junin Quechua* of South America, *speak a language devoid of mental-state terms and yet have a rich narrative tradition*. When retelling Junin Quechua folk tales in their own words, Western people seem compelled to insert terms denoting specific mental states where there were none in the original. Also, Junin Quechua children, despite having no native word for "belief" *perform as well as their western counterparts on so-called false belief tasks*.

Hutto's **Narrative Practice Hypothesis** (Hutto 2008b, 2008a) proposes that a facility with narratives is the key to the abilities usually incorporated in folk psychology. I suggest that although the correlation is interesting an alternative explanation is that *facility with narratives underpins belief-desire psychology as the dominant narrative of action within our culture*.

I speculate that many of the features of the way that we process narratives and the tendency of dominant narratives to produce systematic biases suggest that we might employ a **narrative heuristic**: this would require further empirical work to confirm.

The principal arguments of Chapter Four are:

1) People tend to describe the actions of agents – or apparent agents – by means of narratives.

2) Narrative is an essential component of the way that people structure their experience of the world.
3) Narrative psychology shows that constructing a narrative can have significant effects on the way we process experience. The predominant stories that we construct shape our understanding of events including our own actions.
4) Belief-desire psychology is one particular way to describe action but it is not essential: the same action can be described without them.
5) Belief-desire psychology is a cultural artefact – other cultures describe action without the use of similar terms.
6) The way that we assess the acceptability of a narrative shares features with heuristic judgements: it is conceivable that we judge the meaningfulness of narratives heuristically rather than against historical correspondences or future probabilities.

Chapter Five developed an area in which claims of specific beliefs and desires feature more prominently – the offering of **excuses**. Since the giving of a **justification** for an anomalous, morally suspect or illegal action depends on establishing some feature of the act that justifies it, rather than a feature of the actor, *any time an individual claims a particular belief and or desire in an attempt to explain their behaviour or to avoid moral opprobrium or legal censure they are, in fact, offering an excuse*. In developing this distinction, I examine the epistemology of testimony as applied to justifications and excuses and ask *whether the belief-desire component of those excuses that feature such terms is intended* to be **assessed** or **evaluated** as are ordinary cases of giving **testimony**. I conclude that excuse-givers and their audiences are sensitive to the distinction that the *objective of an excuse* is not testimony as to an individual's beliefs and desires, nor to establish the causal link between specific states and the subsequent behaviour. *The objective of an excuse is to persuade that the behaviour under examination was reasonable under the circumstances*.

This contention was made more specific by an examination of the legal concept of ***mens rea*** (guilty mind) – a component of many legal prohibitions (or requirements). *Mens rea* demands that *the intentions of the accused form part of the definition of those offences* of which it is a component. I point out that the main purpose of the inclusion of a *mens rea* component in the formulation of a law is to allow defences based on the absence of a guilty mind, where the prosecution fails to establish *beyond a reasonable doubt* that the defendant

acted *intentionally*.[102] Once again the standard is one of reasonableness – and reasonableness depends much more on what is at stake, on the degree of risk, than on the specific propositional attitudes of the alleged miscreant.

Two case histories illustrate this point and the importance of overall circumstances to an understanding of reasonableness. The first, R v Morgan et al, led to a change to the law when it was realised that *the offence of rape, as previously formulated, allowed a defence based on what the defendant "genuinely believed"* at the time of the offence. There was, in short, no room for the jury to decide whether the belief or the subsequent actions were reasonable. In its place, most recent formulation of the law specifies the defendant in a rape trial who relies on a *mens rea* defence that they thought the act was consensual must show that they had *taken all reasonable steps to secure consent*. The second case was a notorious shooting of a policeman in London in the 1950s. The older accomplice of the shooter was hanged: the shooter himself, although convicted, was too young to suffer the death penalty. The guilt of the accomplice rested on *two questions pertinent to his knowledge and his intentions – did he know that the younger man had a gun and was he party to the decision to use it*? Forty-six years later, the conviction was posthumously quashed *because the jury had not been directed to base their estimation of his intentions on all the salient facts about his behaviour* at the time of the shooting. The law thus recognised that *the attribution of intention depends on a holistic view of the circumstances under which the action is performed and a judgement of the reasonableness of the action under those circumstances*.

The principal arguments of Chapter Five are, therefore:

1) That **excuses** couched in terms of beliefs and desires are *judged on the basis of what is reasonable under the circumstances*.
2) When laws are formulated with a ***mens rea*** (guilty mind) component, the objective is to *exclude unintentional actions* from liability under the law rather than to specify which mental states imply guilt.
3) Bad laws and unintended legal consequences result when metaphysical commitments similar to those of belief-desire psychology creep into the courtroom. Steps have been taken specifically to exclude mental-state ascriptions

[102] As distinct from "unintentionally" – no ascription of specific mental states is entailed in this definition.

> from the exercise of justice without due regard to the *circumstances* under which the accused acted.

In **Chapter 6** I describe another occasion on which we might offer a statement that, superficially at least, makes direct reference to the possession of a specific belief. **Hedging** is the name given to the phenomenon whereby we seek to modify a statement in order to make it *more or less "fuzzy"*. There are a number of reasons why we might want to do so. In this chapter I draw attention to three possible functions of hedging a statement: **approximation**, **shielding** (avoiding censure if the claim should prove to be incorrect) and **politeness** (in order to avoid a charge of arrogance or disrespect). In all three cases, hedging is a category of speech act in which we *either modify or appear to modify our degree of epistemic commitment to a proposition*. Preceding the statement of a proposition with "I believe that" or inserting a parenthetic "I believe" into an otherwise **constative** (Austin 1962) utterance serves the purpose of giving the impression that one cannot claim to know that it is true: it is less definitive than the **assertion** of the proposition.

I discuss the **conversational implicatures** (Grice 1991; Levinson 2001) of hedges and specifically of "I believe that" as a hedging phrase. I suggest that this suggests a further divergence from the metaphysical commitments of the belief-desire picture.

I highlight the way that we understand "I believe that", used as a hedge, to be modifying the speaker's epistemic commitment to the attached proposition, in this usage the phrase is being used to refer to an epistemic state – the less committed "believe" rather than "know". However, in the case of "I believe that", used as a hedging statement, need not be regarded as the *assertion of an additional causal fact* in order to be understood as picking out an epistemic state of the speaker. Its elimination from accounts of the *causes of action* would not preclude its performance of this linguistic function even if the word "believe", in this context, retains a very ordinary meaning.

Arguments from Chapter 6:

1) Hedges serve to make an asserted statement more or less fuzzy.
2) Hedges also perform a number of communicative functions, including approximation, shielding the speaker and rendering a statement more polite.
3) "I believe that…" can serve as a hedging phrase and can perform any of these functions.

4) In all cases, the addition of "I believe that" either modifies or appears to modify the speaker's epistemic commitment to the attached proposition.
5) Despite "I believe that", used as a hedging phrase, retaining the meaning of designating a particular category of mental state, its elimination from causal accounts of action modest (eliminativism) would not affect its ability to perform this role and so remains a live possibility.

7.3 Further Investigations

In conclusion I would like to suggest some further avenues for philosophical and psychological research suggested by the discussions in this thesis.

In Chapter 4 I suggest that, given the ubiquity and apparent psychological significance of narratives, it is possible that *some of our evaluations rest not on the truth of what is said or written but on our evaluation of its merits as a story*. Such evaluations might rest on *how good a story* we think it is, a judgement which itself might depend on a comparison with the content and structure of story archetypes that prevail in our culture. Is there, I asked, a **narrative heuristic**? I suggest that such a tool might depend on simple facilities like *recognition* and *fluency* yet might combine these in ways that are uniquely suited to the evaluation of stories. I also suggested the kinds of rules that might make up a narrative heuristic.

It would be an interesting research project to take up the investigation of this speculation and to uncover whether such a heuristic exists and, if it does, in what circumstances and with what consequences it is employed. The evaluation of a heuristic depends on two factors:

i) Establishing the features of a heuristic - its cues, search rule, stopping rule and decision rule, for example, which might be investigated by computer simulation (Gigerenzer and Todd 1999a; Gigerenzer and Gaissmaier 2011).
ii) Determining whether the heuristic is actually used in the real world, which can only be determined by empirical investigation of real people (Gigerenzer and Brighton 2011).

Applying these and the techniques for modelling, testing and investigating heuristics "in the wild" should be used to establish both how a narrative heuristic might work and whether it is a decision strategy that people use.

7.4 Embodied and Enactive Cognition.

Beginning in the 1990s, some psychologists and philosophers alike started to express dissatisfaction with the reductive ambitions of some **computational models of mind**. Varela et al. (1993) suggested that there must be a point of contact between the scientific model of the mind and the way that individuals experience and engage with their world. Lakoff and Johnson (1999) develop the idea of an embodied cognition from the idea that so many of our abstract concepts are developed by means of **metaphorical extension** from our physical engagement with the world – "grasping an idea," for example. Their wide-ranging treatment considers the implications of an embodied approach for all kinds of traditional philosophical concerns, from the metaphysics of causation to morality, from conceptions of self to notions of truth and reality.

E. Thompson (2007) seeks to take enactive cognition further and to reconcile its consequences with both phenomenological philosophy and the biology of neuronal systems. However, this ambitious project makes clear that one does not have to embrace all of the methods or conclusions of phenomenology to find this way of proceeding valuable. All that he proposes is the re-integration of the notion of *experience* into the models of the mind that we construct. As he writes:

> The computer model or computational theory of mind, once considered 'the only game in town,' is now called classical cognitive science and coexists, separately and in various hybrid forms, with connectionism and embodied dynamicism. Consciousness, once dismissed as marginal to the scientific understanding of the mind, is now a subject of great interest.
>
> (Ibid.: 267)

This suggests that we might reduce the processes of thought, decision making or interpersonal understanding to symbol manipulations but should treat *meaning* as significant in and of itself. Seeing all of these processes as depending on the manipulation of symbols that stand for "beliefs" and "desires" abstracts away from such direct considerations.

In a collection edited by Zlatev et al. (2008) several authors suggest that the paradigmatic way that humans come to develop their cognition of the world and their specific conceptualisations is by engagement with other people (Gallagher and Hutto 2008; Hobson

and Hobson 2008; Susswein and Racine 2008). These **intersubjective approaches** take it as fundamental to our "mental" life not only that we must negotiate the world of experience but also that, through communication and interaction with others, that much of what we process is shared (cf. Butterfill 2013). Susswein and Racine (2008) draw on the concept of **speech acts** from Austin (1962) to explain the source of confusion that "psychological predicates … are used in a variety of relatively unrelated ways. When subsuming mental state concepts under the superordinate category of 'the mind' it is easy not to notice this feature" (Susswein and Racine 2008: 150). It has been the contention of the present thesis – particularly in Part Two – that philosophers must be sensitive to these differences and to the fact that "a single mental state term can be manifestly used to perform very different social acts" (Ibid.: 151).

Subtitled "Basic Minds without Concepts", Hutto and Myin (2013) present a radical conception of the role of enactive engagement with the world to the development of what was formerly regarded as a distinct realm of "the mental". They criticise the "intellectualists" for whom "nothing qualifies as an action proper unless it is produced by or otherwise connected to contentful states of mind of some sort" (Ibid.: 14). In place of this model (of which belief-desire psychology is a subspecies) they propose an embodiment thesis that "equates basic cognition with concrete spatio-temporally extended patterns of dynamic interaction between organisms and their environments" (Ibid.: 5). Such embodiment "is not defined with reference to an intuitive, everyday understanding of bodies and their boundaries, but in terms of wide-reaching sensorimotor interactions that are contextually embedded" (Ibid.: 6). A consequence is that "not all mentality requires individuals to construct representations of their worlds" (Ibid.: 5). The proponent of a traditional belief-desire psychology would have to reject such a position out of hand.

The rejection of the default idea that cognition consists in computational operations carried out over abstract symbols removed from the experiences that they stand for is a common feature of many *enactive, intersubjective and embodied approaches*. In its place, these accounts suggest that experiences are themselves the currency of cognition. *Their manipulation takes place in the imagination, which is not a separate space from the space of direct perceptual experience*.

Shapiro (2011) divides the embodied approach to three areas of concern: **conceptualisation**, which posits that the embodied nature of the human constrain how we

conceive the world; **replacement**, under which the traditional investigations of cognitive science can be rejected wholesale in favour of an embodied approach and finally theories of **constitution**, which suggests that embodiment and the organism's engagement with the world should be regarded as constituting its mental life, and so abandons the "inner-outer" dichotomy altogether. He suggests that conceptualisation competes unsuccessfully with standard *cognitive science*, replacement directly competes with more success and constitution approaches do not compete, instead treating the explanatory task in an altogether different way (Ibid.: 210).

Embodied and enactive approaches offer an alternative way to view the mind. It eschews the notion of mental functions as something computationally distinct from an organism's engagement with its environment in favour of taking into account all of the factors which make it – an agent or a person – able to operate in the world. It offers the opportunity to develop holistic approaches which are more able to encompass complexity and nuance than computationally reductive schemata.

7.5 A Different "Rationality"

Philosophers have developed normative prescriptions for judgement and decision making. The result has been the formulation of a number of *idealised accounts of rationality*. We are expected to judge a person as rational according to how closely their decision-making process resembles the ideal. We appraise a particular judgement or decision, similarly, according to whether the process used to arrive at it.

A move to more ecological accounts of rationality suggests that we should, instead, judge people and their decisions according to the appropriateness of the outcomes of judgements and decisions. This entails taking into account the environmental, situational and embodied and enactive features of humans. This is **ecological rationality**, the model for which Simon (1991) illustrated with the metaphor of a pair of scissors, in which one blade of the decision is provided by the cognitive processes of the decider and the other by the environmental, situational and consequential constraints and requirements under which the decision takes place (see Chapter 2, Section 2.5 for more on ecological rationality).

Researchers have developed sophisticated models of the **heuristic rules** that people use to solve complex reasoning and social problems in the real world. The collections edited by

Gigerenzer and Todd (1999b) and by Gigerenzer, Hertwig and Pachur (2011) provide an overview of the scope of this field.

The adoption of ecological rationality does not imply modifying the normative in order to encompass the descriptive, "deriving an ought from an is", so to speak. It does suggest that our normative model of something as important as rationality should be sensitive to human and environmental constraints. *The normative ideal should be feasible, regardless of whether or not it is actual*.

Two strategies are available, but we should ask which is used to solve the problems in real-world situations. Even people at the socially impaired end of the autistic spectrum can be taught to apply belief-desire reasoning to pass false-belief tasks and can use this to negotiate the social world (Begeer et al. 2011). This does not mean that *this is how neurotypical people work most of the time*.

It is just this pairing of strategies that has prompted the development of dual-process or dual-system models of judgement and decision making, although some – for example Keren and Schul (2009) – question whether we need to posit separate systems or to regard the strategies that people use as on a continuum (see section 2.10). Wherever the next step in the empirical investigation of judgement, decision making and the psychological antecedents of action leads, the central contentions of this thesis remain:

a) Belief-desire psychology does not explain the causal origins of actions.
b) Belief-desire psychology does not capture the way that people usually account for their own or other people's actions.
c) On those limited occasions that the terms "belief" and "desire" are deployed they should not necessarily be understood as *causal explanations of actions*.

Philosophers interested in human rationality and interpersonal understanding should be prepared to embrace this growing weight of empirical evidence that the strategies used in these fields owe more to heuristics and to automatic processes than is allowed for by the traditional view of rationality. The belief-desire model is an attempt to shoehorn the computational, information-processing model of the mind and of thought into a traditional framework. New models are required.

This final chapter begins with a quotation from Alexander Pope that prefigures that from Wittgenstein that headed the first. If we want to know **why people do what they do**, to

contemplate other minds, and to **know ourselves,** we need to pay attention to people, in their entirety. Philosophers are well placed to draw on the empirical findings from a variety of disciplines – psychology, anthropology, sociology, anatomy, linguistics and others – to elucidate how meaning emerges from the totality of human experience. This synthesis is a fruitful potential area for philosophical investigation. Science cannot answer essentially philosophical questions. An engagement with science can, however, inform philosophical consideration and send the enquiring philosopher off into exciting new directions.

Rationality has been a central concern of Western philosophy at least since the Classical Greek period. Not only in the sense that we must regard ourselves as rational beings if normative ethical theories are to have any weight (Scanlon 2014), but also because if we are to judge our own behaviour and the behaviour of other intentional agents (human and non-human), then we must have an idea of what it means to be "rational".

My appeal is that philosophers shake off the shackles of the belief-desire law and embrace new models of rationality. In some measure these are heuristic, to be sure, automatic, certainly, possibly embodied and potentially without the conundrum of "mental representation" at its centre. None of this can occur unless we free ourselves from the *unwarranted picture that causal explanations of action ought to advert to the specific referents of "belief" and "desire" or similar propositional attitudes*.

7.6 Experimental Philosophy

Philosophers need not leave all of the empirical work on these new concepts of human rationality to psychologists. In recent years, some philosophers have started to apply the techniques and tools of social science and psychology to shed new light on philosophical questions. This approach is especially apt in fields where the traditional answer has rested on a particular, presumably general, intuition.

Describing intentional action has been a subject for experimental philosophy since its beginning. Perhaps the most famous and oft-replicated philosophical experiment is that in which the "Knobe effect" first appeared. Joshua Knobe investigated subjects' response to the following vignette.

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'

> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> (Knobe 2003, 2006)

Subjects were asked whether or not the chairman of the board *intentionally harmed the environment*. Most people (82%) asked answered that they thought that he did (Knobe 2003, 2006). Other subjects were offered this different scenario.

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
>
> The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> (Ibid.)

When asked whether the chairman of the board *intentionally helped the environment*, most people (77%) who considered this scenario said that he did not (Knobe 2003, 2006*)*. Yet in both versions the *chairman's declared intention is exactly the same* – to make as much profit as he can. Knobe's finding, that our moral evaluation of the chairman's action seems to affect our estimate of the degree to which he *intended* to harm/help the environment is the effect that now bears his name. One might speculate that we are *unwilling to give credit (responsibility) to somebody who does the right thing incidentally to the pursuit of their overriding goals*. It is also consistent with the idea that our sense of culpability/responsibility is dependent (at least in part) on our assessment of what is at stake (see Chapter 5, section 5.3).

Knobe's overall conclusion is that *despite the appeal of the folk-psychology picture of our understanding of intentional action, there is something "not quite right" about it* (Knobe 2006: 204).

For some, of course, the idea that philosophers might embark on their own empirical investigations is anathema. In my own, modest view, it is a source of data which *provides philosophical reflection with a perspicuous representation of its target phenomena*. This is

also my justification for philosophical engagement with empirical science. It is also a way that intuitions, long a presumed source of philosophical insight, can be tested, systematically, by controlling variables (as in the Chairman example) and discovering what effect small changes have on our own intuitions and those of others (Fischer 2014; Fischer et al. 2015).

Experimental approaches could provide additional data about the target phenomenon of the present thesis – how people understand and describe their own and each other's actions. Matthew Ratcliffe recounts an informal study that he carried out among a group of students taking a second-year "philosophy of mind" module. He set them this question:

> What is central to your understanding of others? To put it another way, understanding or interacting with another person is very different from understanding or interacting with a rock. What does that difference consist of? Please state your intuitive or commonsense view rather than stating philosophical positions or engaging in philosophical argument.
>
> Ratcliffe (2007: 46)

Ratcliffe reports that his students "mentioned a diverse range of factors". Of the 25 students polled, "The term 'belief' appeared twice in total and 'desire' only appeared once, as did 'prediction'. 'Explanation' was not mentioned at all" (Ibid.: 48). Instead, the kinds of strategy that the students listed included such items as "Can detect their emotions through facial expressions and body language," "Empathy," and "They act similarly to us,". Ratcliffe reports repeating the exercise two years later (2005) and receiving similar responses.

Data from a larger and more diverse sample group, using a more systematic questionnaire to test the intuitions that people really have about their sense of other people would help to inform the debate. Results from different groups should be compared to see what effect philosophical (and psychological) training has on these intuitions. This is only one of a variety of experimental approaches that might be valuable in shedding further empirical light directly on these issues.

7.7 Conclusions

From the arguments developed in each of these chapters I would suggest that we can draw these conclusions:

I) Many commitments of the belief-desire psychology picture are unwarranted by the lights of the scientific understanding of the phenomena (action, interpersonal, understanding and reason-giving) and by the way that their terms ("belief" and "desire") are used in everyday discourse.

II) Many of the philosophical problems generated by the belief-desire picture are artefacts of that fixed way of looking at things, and of metaphysical uses of "belief" and "desire".

III) We can loosen the grip of the assumptions of philosophical folk psychology by taking a more inclusive overview of the phenomena, embracing cognitive and social psychology, and allowing for a wider understanding of the way that these terms are used in everyday situations. This is what Wittgenstein describes at *Philosophical Investigations* §122 as a **perspicuous representation** *(übersichtliche darstellung)* of the phenomena: such a view allows us to take account of the network of connections and the significance of intermediate cases.

IV) Liberation from the assumptions of the belief-desire picture is facilitated by the elimination of "belief" and "desire" from *philosophically rigorous accounts of the causes of action*.

V) If the terms "belief" and "desire" are eliminated from generalised causal accounts of action, *philosophical problems* generated by that picture dissolve.

I have described my contention that "belief" and "desire" might be eliminated from causal accounts of human action as **modest eliminativism**. This differs from the eliminative materialism of Churchland (1981) and others. It entails only that "belief" and "desire" should be eliminated from *causal* accounts of action.

Paradoxically, a causal account of action is both more complicated and more simple than the belief-desire picture suggests. More simple if we want a satisfactory answer to the question "why did you act as you did?" More complicated if we demand a scientifically satisfactory causal-explanatory account of the factors that led to any individual carrying out any particular action.

A robust philosophical account of action should both respect the scientific evidence of the causes of action and be able to accommodate individual and cultural differences in sensitivity to influences, decision strategy, choice and behaviour. Generalisations such as the belief-desire law *might not even serve as a normative account* – since even ideals should, in principle, be within the range of what is possible for a human being.

Subsuming the gamut of causally efficacious conditions under "belief" and "desire" gives rise to philosophical problems: philosophers then expend valuable time and intellectual energy in pursuit of solutions. This behaviour is, I contend, *an indication for application of a diagnostic-therapeutic philosophy.*

Applying such an approach suggests that a **modest elimination** is called for.

8 Bibliography and References

Aarts, Bas, et al. (2004), *Fuzzy Grammar: a reader* (Oxford: Oxford University Press).

Abbott, H. Porter (2007), 'Story, plot and narration', in David Herman (ed.), *The Cambridge Companion to Narrative* (Cambridge: Cambridge University Press), 39-51.

--- (2008), *The Cambridge Introduction to Narrative* (Second edn.; Cambridge: Cambridge University Press).

Alexander, Larry and Kessler Ferzan, Kimberly (2009), *Crime and Culpability: a Theory of Criminal Law* (Cambridge: Cambridge University Press).

Alter, A. L., et al. (2007), 'Overcoming Intuition: Metacognitive difficulty activates analytic reasoning', *Journal of Experimental Psychology: General,* 136 (4), 569-76.

Alvarez, Maria (2010), *Kinds of Reasons: An Essay in the Philosophy of Action* (Oxford: Oxford University Press).

American-Law-Institute (1985), 'Model Penal Code and Commentaries', (Philadelphia).

Anderson, John R. (1995), 'ACT: A simple model of complex cognition', *American Psychologist,* 51 (4), 355-65.

Antony, Louise (2015), 'Defending Folk Psychology', in Robert N Johnson and Michael Smith (eds.), *Passions and Projections: Themes from the Philosophy of Simon Blackburn* (Oxford: Oxford University Press), 3-24.

Apperley, Ian A. and Butterfill, Stephen A. (2009), 'Do Humans Have Two Systems to Track Beliefs and Belief-Like States', *Psychological Review,* 116 (4), 953-70.

Apperly, Ian A. (2011), *Mindreaders: The Cognitive Basis of Theory of Mind* (Hove: Psychology Press).

Armstrong, D. M. (1988), 'The Causal Theory of the Mind', in William G Lycan and Jesse J Prinz (eds.), *Mind and Cognition: An Anthology* (Oxford: Blackwell), 31-47.

Austin, John Langshaw (1962), *How to do things with words* (Oxford: Clarendon Press).

--- (1979a), 'A Plea for Excuses', in J O Urmson and Geoffrey James Warnock (eds.), *J L Austin: Philosophical Papers* (Third edn.; Oxford: Clarendon/Oxford University Press), 175-204.

--- (1979b), 'Other Minds', in J O Urmson and Geoffrey James Warnock (eds.), *J L Austin: Philosophical Papers* (Third edn.; Oxford: Clarendon), 76-116.

--- (1979c), 'Three Ways of Spilling Ink', in J O Urmson and Geoffrey James Warnock (eds.), *JL Austin: Philosophical Papers* (Oxford: Clarendon), 272-87.

Baddeley, Alan, Eysenck, Michael, and Anderson, Michael C. (2009), *Memory* (Hove: Psychology Press).

Baker, Gordon (2004), *Wittgenstein's Method: Neglected Aspects* (Oxford: Blackwell).

Baker, Gordon and Hacker, P. M. S. (2009), *Wittgenstein: Understanding and Meaning* (2nd Revised edn., An Analytic Commentary on the Philosophical Investigstions, 1; Oxford: Wiley-Blackwell).

--- (2014), *Wittgenstein: Rules, Grammar and Necessity* (An Analytic Commentary on the Philosophical Investigations, 2; Oxford: Wiley-Blackwell).

Baldwin, Dare A. (2000), 'Interpersonal Understanding Fuels Knowledge Acquisition', *Current Directions in Psychological Science,* 9 (2), 40-45.

Baldwin, J. F. (1979), 'Fuzzy Logic and Fuzzy Reasoning', *International Journal of Man-Machine Studies,* 11, 465-80.

Bargh, John A. (1994), 'The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition', in R S Wyer Jr and T K Srull (eds.),

Handbook of Social Cognition (2nd edn.; Hillsdale NJ: Laurence Erlbaum and Associates), 1-40.
Bargh, John A. and Chartrand, Tanya L. (1999), 'The Unbearable Automaticity of Being', *American Psychologist,* 54 (7), 462-79.
Bargh, John A. and Ferguson, Melissa J. (2000), 'Beyond Behaviorism: On the Automaticity of Higher Mental Processes', *Psychological Bulletin,* 126 (6), 925-45.
Bargh, John A., Chen, Mark, and Burrows, Lara (1996), 'Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action', *Journal of Personality and Social Psychology,* 71 (2), 230-44.
Baron-Cohen, Simon (1995), *Mindblindness: An Essay on Autism and Theory of Mind* (Cambridge Ma & London: Bradford/MIT press).
Baron-Cohen, Simon, Leslie, Alan M, and Frith, Uta (1985), 'Does the autistic child have a "theory of mind"?', *Cognition,* 21, 37-46.
Barthes, Roland (1988), *Introduction to the structural analysis of the narrative* (University of Birmingham, Centre for Contemporary Cultural Studies).
Bechtel, William and Wright, Cory D. (2009), 'What is Psychological Explanation?', in John Symons and Paco Calvo (eds.), *The Routledge Companion to Philosophy of Psychology* (Oxford: Routledge), 114-30.
Beck, Aaron T. (1979), *Cognitive Therapy and the Emotional Disorders* (New York, London: Meridian/Penguin).
Beck, Judith S. (1995), *Cognitive Therapy; Basics and Beyond* (New York: Guilford Press).
Begeer, Sander, et al. (2011), 'Theory of Mind Training in Children with Autism: A Randomized Controlled Trial', *Journal of Autism and Developmental Disorders,* 41, 997-1006.
Bell, Andy (2002), *Debates in Psychology* (Hove: Routledge).
Bermúdez, José Luis (2005), *Philosophy of psychology : a contemporary introduction* (Routledge contemporary introductions to philosophy; Abingdon & New York: Routledge).
--- (2009), *Decision Theory and Rationality* (Oxford: Oxford University Press).
--- (2010), *Cognitive Science: An Introduction to the Science of the Mind* (Cambridge: Cambridge University Press).
Bingham of Cornhill, Lord, Kennedy, Lord Justice, and Collins, Mr Justice (1998), 'R v Derek William Bentley (Deceased)', in The Supreme Court of Judicature (UK) Court of Appeal (Criminal Division) (ed.), *97/7533/S1* (London).
Birnbaum, Michael H. (2004), 'Base Rates in Bayesian Inference', in Rüdiger F Pohl (ed.), *Cognitive Illusions* (Hove: Psychology Press), 43-60.
Bjork, Robert A. (1989), 'Retrieval Inhibition as an Adaptive Mechanism in Human Memory', in Henry L. Roediger and Fergus I. M. Craik (eds.), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving* (Hillsdale NJ: Lawrence Erlbaum Assoc.), 309-30.
Blass, Thomas (1999), 'The Milgram Paradigm after 35 years: Some Things We Now Know About Obedience to Authority', *Journal of Applied Social Psychology,* 29 (5), 955-78.
Block, Ned (2004), 'What is Functionalism?', in John Heil (ed.), *Philosophy of Mind, a guide and anthology* (Oxford: Oxford University Press), 183-99.
Bloor, Thomas and Bloor, Meriel (2004), *The Functional Analysis of English* (London: Arnold).

Bodenhausen, Galen V and Todd, Andrew R (2010), 'Automatic Aspects of Judgement and Decision Making', in Bertram Gawronski and B Keith Payne (eds.), *Handbook of Implicit Social Cognition* (London & New York: Guildford Press), 278-94.
Borges, Bernhard, et al. (1999), 'Can ignorance beat the stock market?', in Gerd Gigerenzer and Peter M. Todd (eds.), *Simple Heuristics that Make us Smart* (Oxford: Oxford University Press), 59-72.
Bortolotti, Lisa (2010), *Delusions and Other Irrational Beliefs* (Oxford Oxford University Press).
Brandom, Robert (1994), *Making it Explicit: Reasoning, Representing and Discursive Commitment* (Cambridge MA: Harvard University Press).
Brewin, Chris R. (2001), 'memory Processes in Post-Traumatic Stress Disorder', *International Review of Psychiatry,* 13, 159-63.
Brockner, Joel, Rubin, Jeffrey Z., and Lang, Elaine (1981), 'Face Saving and Entrapment', *Journal of Experimental Social Psychology,* 17, 68-79.
Broome, John (2013), *Rationality Through Reasoning* (Chichester: Wiley &Sons).
Brown, Jessica and Cappelen, Herman (2014), 'Assertion: An Introduction and Overview', in Jessica Brown and Herman Cappelen (eds.), *Assertion: New Philosophical Essays* (Paperback edn.; Oxford: Oxford University Press), 1-17.
Bruner, Jerome S. (1990), *Acts of meaning* (Cambridge MA: Harvard University Press).
--- (1991), 'The narrative construction of reality', *Critical inquiry,* 18 (1), 1-21.
Burge, Tyler (1993), 'Content Preservation', *The Philosophical Review,* 102 (4), 457-88.
Burns, Bruce D. and Wieth, Mareike (2004), 'The collider principle in causal reasoning: why the Monty Hall dilemma is so hard', *Journal of Experimental Psychology: General,* 133 (3), 434.
Butterfill, Stephen (2013), 'Interacting Mindreaders', *Philosophical Studies,* 165, 841-163.
Butterfill, Stephen and Apperley, Ian A (2012), 'How to Construct a Minimal Theory of Mind', *Mind and Language,* 28 (5), 606-37.
Call, Josep and Tomasello, Michael (1999), 'A Nonverbal False Belief Task: The Performance of Children and Great Apes', *Child Development,* 70 (2), 381-95.
Camerer, Colin (1995), 'Individual decision making', in J H Kagel and A E Roth (eds.), *Handbook of Experimental Economics* (Princeton NJ: Princeton University Press), 587-703.
Campbell, John (2009), 'What does rationality have to do with psychological causation? Propositional attitudes as mechanisms and as control variables.', in Matthew R Broome and Lisa Bortolotti (eds.), *Psychiatry as Cognitive Neuroscience* (Oxford: Oxford University Press), 137-49.
Carnahan, Thomas and McFarland, Sam (2007), 'Revisiting the Stanford Prison Experiment: Could Participant Self-Selection Have Led to the Cruelty?', *Personality and Social Psychology Bulletin,* 33 (5), 603-14.
Carr, David (1986), *Time, Narrative and History* (Bloomington: Indiana University Press).
Carruthers, Peter (2009), 'How we know our own minds: The relationship between mindreading and metacognition', *Behavioural and Brain Sciences,* 32, 121-82.
Carson, David C. and Felthous, Alan R. (2003), 'Mens Rea', *Behavioural Sciences and the Law,* 21, 559-62.
Chaiken, Shelly (1980), 'Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion', *Journal of Personality and Social Psychology,* 39 (5), 752-66.

Chapman, Gretchen B. (2004), 'The Psychology of Medical Decision Making', in Derek J. Koehler and Nigel Harvey (eds.), *Blackwell Handbook of Judgement and Decision Making* (Oxford: Blackwell), 585-602.
Chen, Serena, Fitzsimmons, Gràinne M., and Andersen, Susan M. (2007), 'Automaticity in Close Relationships', in John A Bargh (ed.), *Social Psychology and the Unconscious; The Automaticity of Higher Mental Processes* (New York & Hove Psychology Press/Taylor & Francis), 133-72.
Chung, Man Cheung and Hyland, Michael E. (2012), *History and Philosophy of Psychology* (Oxford: Wiley-Blackwell).
Churchland, Paul M. (1981), 'Eliminative Materialism and the Propositional Attitudes', *The Journal of Philosophy,* 78 (2), 67-90.
--- (1988), *Matter and consciousness : a contemporary introduction to the philosophy of mind* (Rev. edn.; Cambridge, Mass. ; London: MIT).
Clemen, Gudrun (1997), 'The Concept of Hedging: Origins, Approaches and Definitions', in Raija Markkannen and Hartmut Schröder (eds.), *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts* (Berlin: De Gruyter), 235-48.
Coady, C. A. J. (1992), *Testimony: A Philosophical Study* (Oxford: Clarendon Press).
Collins, Allan M. and Quillian, M. Ross (1969), 'Retrieval Time from Semantic Memory', *Journal of Verbal Learning and Verbal Behaviour,* 8, 240-47.
Collins, Allan M. and Loftus, Elizabeth F. (1975), 'A Spreading Activation Theory of Semantic Processing', *Psychological Review,* 82 (6), 407-28.
Collins, Barry E. and Ma, Laura (2000), 'Impression Management and Identity Construction in the Milgram Social System', *Obedience to Authority: Current Perspectives on the Milgram Paradigm* (Mahwah NJ, London: Lawrence Erlbaum), 61-90.
Conrad, Carol (1972), 'Cognitive economy in semantic memory', *Journal of Experimental Psychology,* 92, 149-54.
Crompton, Peter (1997), 'Hedging in Academic Writing: Some Theoretic Problems', *English for Specific Purposes,* 16 (4), 271-87.
Crossley, Michele L. (2000), *Introducing Narrative Psychology: Self, Trauma and the Construction of Meaning* (Buckingham: Open University Press).
Cummins, Robert (1980), 'Functional Analysis', *Readings in Philosophy of Psychology* (One; Cambridge MA: Harvard University Press), 185-90.
--- (1983), *The Nature of Psychological Explanation* (Cambridge MA: Bradford/MIT Press).
--- (2006), '"How Does it Work" versus "What Are the Laws": Two Conceptions of Psychological Explanation ', in José Luis Bermúdez (ed.), *Philosophy of Psychology: Contemporary Readings* (Oxford: Routledge), 90-98.
Curley, E. M. (1976), 'Excusing Rape', *Philosophy and Public Affairs,* 5 (4), 355-60.
Currie, Gregory and Sterelny, Kim (2006), 'How to Think about the Modularity of Mind Reading', in José Luis Bermúdez (ed.), *Philosophy of psychology : contemporary readings* (Abingdon; New York: Routledge), 524-38.
Damasio, Antonio (2010), *Self Comes to Mind: Constructing the Conscious Brain* (London: Heinemann).
Davey, Graham (ed.), (2008), *Complete Psychology* (Second edn., Abingdon: Hodder Education).
Davidson, Donald (2001), *Essays on Actions and Events* (Oxford: Clarendon Press).
Dawkins, Richard (1989), *The Selfish Gene* (2nd edn.; Oxford: Oxford University Press).

De Neys, Wim and Verschueren, Niki (2006), 'Working Memory Capacity and a Notorious Brain Teaser', *Experimental Psychology,* 53 (2), 123-31.
Dehaene, Stanislas (2014), *Consciousness and The Brain* (New York: Penguin USA).
Dennett, Daniel C. (1988), 'Precis of "The Intentional Stance"', *Behavioural and Brain Sciences,* 11, 495-546.
--- (1989), *The Intentional Stance* (Cambridge MA: Bradford/MIT press).
--- (1997), 'True Believers: The Intentional Strategy and Why it Works', in Brian Haugeland (ed.), *Mind Design II* (Cambridge MA: MIT Press).
Descartes, René (1988), *Selected Philosophical Writings*, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch (Cambridge: Cambridge University Press).
Dienes, Zoltan (2008), *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference* (Basingstoke: Palgrave Macmillan).
Dijksterhuis, A. (2004), 'Think different: The merits of unconscious thought in preference development and decision making', *Journal of personality and social psychology,* 87, 586-98.
Edwards, J. Ll J. (1955), *Mens Rea in Statutory Offences* (London: Macmillan).
Ellis, Albert (1994), *Reason and Emotion in Psychotherapy* (Revised and Expanded edn.; New York: Lyle Stuart).
Elman, Jeffrey L. (2009), 'On the Meaning of Words and Dinosaur Bones: Lexical Knowledge Without a Lexicon', *Cognitive Science,* 33, 547-82.
Elster, J. (ed.), (1986), *Rational Choice* (Oxford: Blackwell).
Epley, Nicholas, Waytz, Adam, and Cacioppo, John T (2007), 'On Seeing Human: A Three-Factor Theory of Anthropomorphism', *Psychological Review,* 114 (4), 864-86.
Erickson, T. D. and Mattson, M. E. (1981), 'From Words to Meaning: A Semantic Illusion', *Journal of Verbal Learning and Verbal Behaviour,* 20, 540-51.
Evans, Jonathan St. B. T. (2003), 'In two minds: dual-process accounts of reasoning', *Trends in Cognitive Science,* 7 (10), 454-59.
--- (2004), 'Biases in Deductive Reasoning', in Rüdiger F Pohl (ed.), *Cognitive Illusions* (Hove: Psychology Press), 127-44.
--- (2006), 'The heuristic-analytic theory of reasoning: Extension and evaluation', *Psychonomic Bulletin & Review,* 13 (3), 378-95.
--- (2008), 'Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition', *Annual Review of Psychology,* 59, 255-78.
--- (2009a), *In Two Minds: Dual Processes and Beyond* (Oxford: Oxford University Press).
--- (2009b), 'How many dual-process theories do we need? One, two or many?', in Jonathan St B T Evans and Keith Frankish (eds.), *In Two Minds, Dual Processes and Beyond* (Oxford: Oxford University Press), 33-54.
--- (2010), 'Intuition and reasoning: a dual-process perspective', *Psychological Inquiry,* 21, 313-26.
Evans, Jonathan St. B. T. and Lynch, J. S. (1973), 'Matching Bias in the Selection Task', *British Journal of Psychology,* 64 (3), 391-97.
Eysenck, Michael W. and Keane, Mark T. (2010), *Cognitive Psychology; A Student's Handbook* (New York and Hove: Psychology Press).
Ezzy, Douglas (1998), 'Theorizing Narrative Identity; Symbolic Interactionism and Hermeneutics', *The Sociological Quarterly,* 39 (2), 239-52.
Feldman Barrett, Lisa, Ochsner, Kevin N, and Gross, James J (2007), 'On the Automaticity of Emotion', in John A Bargh (ed.), *Social Psychology and the*

Unconscious; The Automaticity of Higher Mental Processes (New York & Hove Psychology Press/Taylor & Francis), 173-217.
Feltzer, Anita (2014), 'I think, I mean and I believe in political discourse', *Functions of Language,* 21 (1), 67-94.
Feyerabend, Paul (1963), 'Materialism and the mind-body problem', *The Review of Metaphysics*, 49-66.
Fischer, Eugen (2006), 'Philosophical Pictures', *Synthese,* 148 (2), 469-501.
--- (2011a), 'How to Practise Philosophy as Therapy: Philosophical Therapy and Therapeutic Philosophy ', *Metaphilosophy,* 42 (1-2), 49-82.
--- (2011b), *Philosophical Delusion and its Therapy* (London, New York: Routledge).
--- (2014), 'Mind the Metaphor! A systematic fallacy in analogical reasoning', *Analysis,* 75 (1), 67-77.
Fischer, Eugen and Engelhardt, Paul E. (2016), 'Intuitions' Linguistic Sources: Stereotypes, Intuitions and Illusions', *Mind and Language,* 31 (1), 67-103.
Fischer, Eugen, Engelhardt, Paul E., and Herbelot, Aurélie (2015), 'Intuitions and Illusions: from explanation and experiment to assessment', in Eugen Fischer and John Collins (eds.), *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method* (Abingdon, New York: Routledge), 259-92.
Fisher, Walter R. (1984), 'Narration as a Human Communication Paradigm: The Case of Public Moral Argument', *Communication Monographs,* 51, 1-22.
Fiske, Susan T. and Taylor, Shellley E. (2013), *Social Cognition* (2nd edn.; London, Thousand Oaks CA: Sage).
Fletcher, Garth (1995), *The Scientific Credibility of Folk Psychology* (Mahwah, N.J.: Lawrence Erlbaum Associates).
Fodor, Jerry A. (1975), *The Language of Thought* (Cambridge MA, London: Harvard University Press).
--- (1987), *Psychosemantics: The problem of meaning in the philosophy of mind* (Cambridge MA, London: Bradford/MIT Press).
--- (1991a), 'Fodor's guide to mental representation: the intelligent auntie's vade mecum', in John D Greenwood (ed.), *The Future of Folk Psychology* (Cambridge: Cambridge University Press), 22-50.
--- (1991b), 'You Can Fool Some of The People All of The Time, Everything Else Being Equal; Hedged Laws and Psychological Explanations', *Mind,* 100 (1), 19-34.
--- (1993), 'The Persistence of the Attitudes', in Scott M Christensen and Dale R Turner (eds.), *Folk Psychology and Philosophy of Mind* (Hillsdale NJ, Hove, London: Laurence Erlbaum), 221-46.
--- (1999), 'Information and Representation', in Eric Margolis and Stephen Laurence (eds.), *Concepts: core readings* (Cambridge MA, London: Bradford/MIT Press), 513-24.
--- (2008), *LOT 2; The Language of Thought Revisited* (Oxford: Oxford University Press).
Försterling, Friedrich (2001), *Attribution: An Introduction to Theories, Research and Applications*, ed. Miles Hewstone (Social Psychology: A Modular Course; Philadelphia PA: Psychology Press).
Fraser, Bruce (2010), 'Pragmatic Competence: The Case of Hedging', in Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider (eds.), *New Approaches to Hedging* (Bingley: Emerald Group), 15-34.
Freud, Sigmund (1999), *The Interpretation of Dreams*, trans. Joyce Crick (Oxford: Oxford University Press).
--- (2003), *An Outline of Psychoanalysis*, trans. Helena Ragg-Kirkby (London: Penguin Classics).

Frye, Northrop (1951), 'The archetypes of literature', *Kenyon Review*, 92-110.
Furnham, Adrian (2003), 'Belief in a just world: research progress over the past decade', *Personality and Individual Differences,* 34 (5), 795-817.
Gallagher, Shaun and Hutto, Daniel D. (2008), 'Understanding others through primary interaction and narrative practice ', in Jordan Zlatev, et al. (eds.), *The Shared Mind: Perspectives on Intersubjectivity* (Amsterdam, Philadelphia PA: John Benjamins), 17-38.
Garcia-Retamero, Rocio and Dhami, Mandeep K. (2011), 'Take-the-Best in Expert-Novice Decision Strategies for Residential Burglary', in Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (eds.), *Heuristics: The Foundations of Adaptive Behaviour* (Oxford: Oxford University Press), 603-09.
Garcia-Retamero, Rocio, Takezawa, Masanori, and Galesic, Mirta (2009), 'Simple Mechanisms for Gathering Social Information', *New Ideas in Psychology,* 28, 49-63.
Gardner, Howard (1987), *The Mind's New Science: A History of the Cognitive Revolution* (New York: Basic Books).
Garfield, Jay L. (1988), *Belief in Psychology: A Study in the Ontology of Mind* (Cambridge MA, London).
Gauker, Christopher (2003), 'The Belief-Desire Law', *Facta Philosophica,* 7, 121-44.
Gazzaniga, Michael S., et al. (2009), *Cognitive Neuroscience* (3rd edn.; New York, London: W W Norton & Company).
Geeraert, Nicolas, et al. (2004), 'The return of dispositionalism: On the linguistic consequences of dispositional suppression', *Journal of Experimental Social Psychology,* 40 (2), 264-72.
Geertz, Clifford (1979), 'From The Native's Point of View: on the nature of anthropological understanding', in Paul Rabinow and William M Sullivan (eds.), *Interpretive Social Science, a reader* (Berkeley: University of California Press).
Gigerenzer, Gerd (1991), 'How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases"', in W. Stroebe and M. Hewstone (eds.), *European Review of Social Psychology* (2; Chichester: John Wiley and Sons), 83-115.
--- (1996), 'On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996)', *Psychological review,* 103 (3), 592.
--- (2008), *Rationality for Mortals: How People Cope with Uncertainty* (Oxford: Oxford University Press).
Gigerenzer, Gerd and Todd, Peter M (1999a), 'Fast and Frugal Heuristics - The Adaptive Toolbox', in Gerd Gigerenzer and Peter M Todd (eds.), *Simple Heuristics That Make Us Smart* (Oxford & New York: Oxford University Press), 3-34.
Gigerenzer, Gerd and Gaissmaier, Wolfgang (2011), 'Heuristic Decision Making', *Annual Review of Psychology,* 62, 451-82.
Gigerenzer, Gerd and Brighton, Harry (2011), 'Homo Heuristicus: Why Biased Minds Make Better Inferences', in Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (eds.), *Heuristics: The Foundations of Adaptive Behaviour* (Oxford: Oxford University Press), 2-27.
Gigerenzer, Gerd and Goldstein, Daniel (2011), 'Reasoning the Fast and Frugal Way: Models of Bounded Rationality', in Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (eds.), *Heuristics: The Foundations of Adaptive Behaviour* (Oxford: Oxford University Press), 31-57.
Gigerenzer, Gerd and Todd, Peter M (eds.) (1999b), *Simple Heuristics that Make us Smart* (Oxford: Oxford University Press).

Gigerenzer, Gerd, Hertwig, Ralph, and Pachur, Thorsten (eds.) (2011), *Heuristics: The Foundations of Adaptive Behaviour* (Oxford: Oxford University Press).
Gilbert, Daniel T and Jones, Edward (1986), 'Perceiver-Induced Constraint: Interpretations of Self-Generated Reality', *Journal of Personality and Social Psychology,* 30 (2), 269-80.
Gilovich, Thomas, Keltner, Dacher, and Nisbett, Richard E (2006), *Social Psychology* (London & New York: W W Norton).
Giora, Rachel (2003), *On Our Mind. Salience, Context and Figurative Language* (Oxford: Oxford University Press).
Glucksberg, Sam, Gildea, Patricia, and Bookin, Howard B. (1982), 'On understanding non-literal speech: can people ignore metaphors?', *Journal of Verbal Learning and Verbal Behaviour,* 21 (1), 85-98.
Godfrey-Smith, Peter (2004), 'On folk psychology and mental representation', *Representation in mind: New approaches to mental representation* (Elsevier), 147-62.
Goldman, Alvin I. (2006), *Simulating Minds; The Philosophy, Psychology and Neuroscience of Mindreading* (Oxford: Oxford University Press).
Goldstein, Daniel and Gigerenzer, Gerd (1999), 'The Recognition Heuristic: How Ignorance Makes us Smart', in Gerd Gigerenzer and Peter M. Todd (eds.), *Simple Heuristics That Make Us Smart* (Oxford: Oxford University Press), 37-58.
Gopnik, Alison and Meltzoff, Andrew N (1997), *Words, Thoughts and Theories* (Cambridge MA & London: Bradford/MIT Press).
Gopnik, Alison and Seiver, Alison (2009), 'Reading Minds: How Infants come to know others', *Zero to Three,* 30 (2), 28-32.
Gordon, Robert M. (2008), 'Folk Psychology as Simulation', in William G Lycan and Jesse Prinz (eds.), *Mind and Cognition, an Anthology* (Oxford: Blackwell), 369-78.
Graham, Peter J. (2000), 'The Reliability of Testimony', *Philosophy and Phenomenological Research,* 61 (3), 695-709.
Greenwood, John D. (1991), 'Introduction: Folk psychology and scientific psychology', in John D Greenwood (ed.), *The Future of Folk Psychology: Intentionality and Cognitive Science* (Cambridge: Cambridge University Press), 1-21.
Grice, Paul (1991), *Studies in the Way of Words* (Paperback edn.; Cambridge MA, London: Harvard University Press).
Hacker, P. M. S. (2007), 'Baker's Late Interpretation of Wittgenstein', in Guy Kahane, Edward Kanterian, and Oskari Kuusela (eds.), *Wittgenstein and his Interpreters* (Oxford: Blackwell), 88-122.
Haney, Craig, Banks, Curtis, and Zimbardo, Philip G. (1973), 'Interpersonal Dynamics in a Simulated Prison', *International Journal of Criminology and Penology* 1, 69-97.
Hannon, Brenda (2014), 'Research on Semantic Illusions Tells us that there are Multiple Sources of Misinformation', in David N. Rapp and Jason L. G. Braasch (eds.), *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and Educational Sciences* (Cambridge MA: MIT Press), 93-116.
Hardman, David (2009), *Judgement and Decision Making* (Oxford: BPS/Blackwell).
Haselager, W. F. G. (1997), *Cognitive Science and Folk Psychology* (London: Sage).
Hassin, Ran R., Bargh, John A., and Uleman, James S. (2002), 'Spontaneous Causal Inferences', *Journal of Experimental Social Psychology,* 38, 515-22.
Hassin, Ran R., Uleman, James S., and Bargh, John A. (eds.) (2005), *The New Unconscious* (Oxford: Oxford University Press).

Heider, Fritz (1958), *The Psychology of Interpersonal Relations* (Hillsdale NJ: Lawrence Erlbaum Associates).
Heider, Fritz and Simmel, Marianne (1944), 'An Experimental Study of Apparent Behavior ', *American Journal of Psychology,* 57, 243-49.
Hempel, Carl G. and Oppenheim, Paul (1948), 'Studies in the Logic of Explanation', *Philosophy of Science,* 15 (2), 135-75.
Herman, David (2007), 'Introduction', in David Herman (ed.), *The Cambridge Companion to Narrative* (Cambridge: Cambridge University Press), 3-21.
--- (2009a), 'Storied Minds', in Daniel D Hutto (ed.), *Narrative and Folk Psychology* (Exeter: Imprint Academic), 40-68.
--- (2009b), *Basic Elements of Narrative* (Oxford: Wiley-Blackwell).
Higgins, E. Tory (2005), 'Motivational Sources of Unintended Thought: Irrational Intrusions or Side Effects of Rational Strategies ', in Ran R. Hassin, James S. Uleman, and John A. Bargh (eds.), *The New Unconscious* (Oxford: Oxford University Press), 516-36.
Hindmoor, Andrew (2006), *Rational Choice* (Basingstoke: Palgrave Macmillan).
Hobson, Peter and Hobson, Jessica A. (2008), 'Engaging, Sharing, Knowing: Some lessons from research in autism ', in Jordan Zlatev, et al. (eds.), *The Shared Mind: Perspectives on Intersubjectivity* (Amsterdam, Philadelphia PA: John Benjamins), 67-88.
Homer (1991), *The Iliad*, trans. R. Fayles (New York: Penguin Classics).
Horgan, Terry and Woodward, James (1991), 'Folk Psychology is Here to Stay', in John D Greenwood (ed.), *The Future of Folk Psychology* (Cambridge: Cambridge University Press).
Horwich, Paul (2012), *Wittgenstein's Metaphilosopy* (Oxford: Clarendon Press).
House, James S., Landis, Karl R., and Debra, Umberson (1988), 'Social Relationships and Health', *Science,* 241 (4865), 540-45.
Huang, Yan (2007), *Pragmatics* (Oxford: Oxford University Press).
Hume, David (1740/1985), *A Treatise on Human Nature* (Penguin Classics).
Hutto, Daniel D. (2008a), 'The Narrative Practice Hypothesis: clarifications and implications', *Philosophical Explorations,* 11 (3), 175-92.
--- (2008b), *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons* (Cambridge MA: Bradford/MIT Press).
--- (2009), 'Folk Psychology as Narrative Practice', in Daniel D Hutto (ed.), *Narrative and Folk Psychology* (Exeter: Imprint Academic), 9-39.
Hutto, Daniel D. and Myin, Eric (2013), *Radicalizing Enactivism; Basic Minds without Content* (Cambridge MA: MIT Press).
Hyland, Ken (1994), 'Hedging in Academic Writing and EAP Textbooks', *English for Specific Purposes,* 13 (3), 239-56.
--- (1995a), 'Getting Serious about Being Tentative: How Scientists Hedge', *New Zealand Studies in Applied Linguistics,* 1, 35-50.
--- (1995b), 'The Author in the Text: Hedging Scientific Writing', *Hong Kong Papers in Linguistics and Language Teaching,* 18, 33-42.
--- (1996), 'Writing without Conviction? Hedging in Science Research Articles', *Applied Linguistics,* 17 (4), 433-54.
--- (1998), *Hedging in Scientific Research Articles* (Amsterdam/Philadelphia: John Benjamins).
Jaynes, Julian (1976), *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (Boston, New York: Mariner (Houghton-Mifflin)).

Johnson, Eric J. and Goldstein, Daniel (2003), 'Do Defaults Save Lives', *Science,* 302, 1338-39.
Johnson, Eric J., et al. (1993), 'Framing, Probability Distortions and Insurance Decisions', *Journal of Risk and Uncertainty,* 7, 35-51.
Johnson, H M (1939), 'Rival Principles of Causal Explanation in Psychology', *The Psychological Review,* 46 (6), 493-516.
Johnson, Kathy E. (2013), 'Culture, Expertise and Mental Categories', in Daniel Reisberg (ed.), *The Oxford Handbook of Cognitive Psychology* (Oxford: Oxford University Press), 330-45.
Johnson, Marcia K, Bush, Julie G, and Mitchell, Karen J (1998), 'Interpersonal Reality Monitoring; Judging the Sources of Other People's Memories', *Social Cognition,* 16 (2), 199-224.
Johnson, Susan C. (2000), 'The recognition of mentalistic agents in infancy', *Trends in Cognitive Science,* 4 (1), 22-28.
Johnson-Laird, Philip (2006), *How We Reason* (Oxford: Oxford University Press).
Jones, Edward E and Nisbett, Richard E (1972), 'The Actor and The Observer: Divergent perceptions of the causes of behaviour', in E E Jones, et al. (eds.), *Atttribution: Perceiving the causes of behaviour* (Morristown NJ: General Learning Press), 79-94.
Kahneman, Daniel (2011), *Thinking, Fast and Slow* (London: Penguin).
Kahneman, Daniel and Tversky, Amos (1972), 'Subjective Probability: A Judgement of Representativeness', *Cognitive Psychology,* 3, 430-54.
--- (1996), 'On the Reality of Cognitive Illusions', *Psychological Review,* 103 (3), 582-91.
Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (eds.) (1982), *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press).
Kaltenböck, Gunther (2009), 'Initial I Think: Main or Comment Clause?', *Discourse and Interaction,* 2 (1), 49-70.
--- (2010), 'Pragmatic Functions of Parenthetical *I Think* ', in Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider (eds.), *New Approaches to Hedging* (Bingley: Emerald Group), 237-66.
Keil, Frank and Newman, George E. (2015), 'Order, Order Everywhere, and Only an Agent to Think: The Cognitive Compulsion to Infer Intentional Agents', *Mind and Language,* 30 (2), 117-39.
Kelley, H H (1973), 'The processes of Causal Attribution', *American Psychologist,* 28, 107-28.
Keren, Gideon and Teigen, Carl H (2004), 'Yet Another Look at the Heuristics and Biases Approach', in Derek J Koehler and Nigel Harvey (eds.), *Blackwell Handbook of Judgment and Decision Making* (Oxford: Blackwell), 89-109.
Keren, Gideon and Schul, Yaacov (2009), 'Two is Not Always Better than One: A Critical Evaluation of Two-System Theories', *Perspectives on Psychological Science,* 4 (6), 533-50.
Kihlstrom, John F. (1988), 'The Cognitive Unconscious', *Science,* 237, 1445-52.
--- (2013), 'Unconscious Processes', in Daniel Reisberg (ed.), *The Oxford Handbook of Cognitive Psychology* (Oxford: Oxford University Press), 176-86.
Kim, Jaegwon (1973), 'Causes and Counterfactuals', *Journal of Philosophy,* 70 (17), 570-72.
--- (2011), *Philosophy of Mind* (Third edn.; Boulder CO: Westview Press/Perseus Books).
Knobe, Joshua (2003), 'Intentional Action in Folk Psychology: an experimental investigation', *Philosophical Psychology,* 16 (2), 309-24.

--- (2006), 'The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology', *Philosophical Studies,* 130, 203-31.
Kruglanski, Arie W. and Gigerenzer, Gerd (2011), 'Intuitive and deliberate judgments are based on common principles', *Psychological Review,* 118 (1), 97-109.
Kuhn, Thomas (1996), *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press).
Kusch, Martin and Lipton, Peter (2002), 'Testimony: A primer', *Studies in History and Philosophy of Science,* 33, 209-17.
Kuusela, Oskari (2008), *The Struggle Against Dogmatism: Wittgenstein and the Concept of Philosophy* (Cambridge MA: Harvard University Press).
Lackey, Jennifer (2006), 'Introduction', in Jennifer Lackey and Ernest Sosa (eds.), *The Epistemology of Testimony* (Oxford: Oxford University Press), 1-21.
Laibson, David (1997), 'Golden Eggs and Hyperbolic Discounting', *Quarterly Journal of Economics,* 112 (2), 443-77.
Lakoff, George (1973), 'Hedges: A study in meaning criteria and the logic of fuzzy concepts', *Journal of philosophical logic,* 2 (4), 458-508.
--- (1989), 'Lexicography and Generative Grammar I: Hedges and Meaning Criteria', *Annals of the New York Academy of Sciences*, 144-53.
Lakoff, George and Johnson, Mark (1999), *Philosophy in the Flesh* (New York: Basic Books/Perseus).
László, János (2008), *The Science of Stories: An introduction to Narrative Psychology* (London: Routledge).
Leech, Geoffrey N (2003), 'Towards an Anatomy of Politeness in Communication', *International Journal of Pragmatics,* 13, 101-23.
Lennon, Alexia, et al. (2011), '"You're a bad driver, but I just made a mistake"; Attribution differences between the "victims" and "perpertrators" of scenario-based aggressive driving incidents.', *Transportation Research Part F,* 14, 209-21.
Levinson, Stephen C (1983), *Pragmatics* (Cambridge: Cambridge University Press).
--- (2001), *Presumptive Meanings* (Cambridge MA: MIT Press).
Lewis, David (1973), 'Counterfactual Dependence and Time's Arrow', *Nous,* 13 (4), 455-76.
--- (1974), 'Radical Interpretation', *Synthese,* 23, 331-44.
--- (1980), 'Psychophysical and Theoretical Identifications', *Readings in Philosophy of Psychology* (1; Cambridge MA: Harvard University Press), 207-15.
--- (2006), 'Reduction of Mind', in José Luis Bermúdez (ed.), *Philosophy of Psychology: contemporary readings* (London, New York: Routledge), 51-63.
Lichtenstein, Sarah and Slovic, Paul (eds.) (2006), *The Construction of Preference* (Cambridge: Cambridge University Press).
Loftus, Elizabeth F (1973), 'Activation of Semantic Memory', *American Journal of Psychology,* 86, 331-37.
Losee, John (2001), *A historical introduction to the Philosophy of Science* (Fourth edn.; Oxford: Oxford University Press).
Louch, A. (1966), *Explanation and Human Action* (Oxford: Basil Blackwell).
MacIntyre, Alisdair (1981), *After Virtue* (London: Bloomsbury).
Maitra, Ishani (2014), 'Assertion, Norms and Games', in Jessica Brown and Herman Cappelen (eds.), *Assertion: New Philosophical Essays* (Oxford: Oxford University Press), 278-95.
Malle, Bertram F (2006), 'The Actor-Observer Asymmetry in Attribution: A (Surprising) Meta-Analysis', *Psychological Bulletin,* 132 (6), 895-919.

Mankiw, N. Gregory and Taylor, Mark P (2014), *Economics* (Andover: Cengage Learning).
Marr, David (2010), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Cambridge MA and London: MIT Press).
Marsh, Barnaby (2002), 'Heuristics as Social Tools', *New Ideas in Psychology,* 20, 49-57.
McAdams, Dan P (1993), *The Stories We Live By: Personal Myths and the Making of the Self* (New York & London: Guilford Press).
McAdams, Dan P and McLean, Kate C (2013), 'Narrative Identity', *Current Directions in Psychological Science,* 22 (3), 233-38.
McCarthy, John and Hayes, Patrick J. (1981), 'Some Philosophical Problems from the Standpoint of Artificial Intelligence', in Bonnie Lynn Webber and Nils J. Nilsson (eds.), *Readings in Artificial Intelligence* (Paperback edn.; Los Altos CA: Morgan Kaufmann), 431-50.
McKee, Robert (1999), *Story* (London: Methuen).
McLeod, Peter and Dienes, Zoltan (1996), 'Do Fielders Know Where to Go to Catch the Ball or Only How to Get There?', *Journal of Experimental Psychology: Human Perception and Performance,* 22 (3), 531-43.
McNamara, T. P. (1992), 'Priming and constraints it places on theories of memory and retrieval', *Psychological Review,* 99, 650-62.
McNeil, Barbara J, et al. (1982), 'On the elicitation of preferences for alternative therapies', *The New England journal of medicine,* 306 (21), 1259.
McRae, Ken and Jones, Michael (2014), 'Semantic Memory', in Daniel Reisberg (ed.), *Oxford Handbook of Cognitive Psychology* (Paperback edn.; Oxford: Oxford University Press), 206-19.
Messer, Stanley B., Sass, Louis A., and Woolfolk, Robert L. (eds.) (1988), *Hermeneutics and Psychological Theory: Interpretive Perspectives on Personality, Psychotherapy and Psychopathology* (Piscataway NJ: Rutgers University Press).
Messick, David (1999), 'Alternative logics for decision making in social settings', *Journal of Economic Behavior & Organisation,* 39, 11-28.
Milgram, Stanley (1974), *Obedience to Authority: an Experimental view* (London: Tavistock).
--- (1994), 'Conformity and Independence', in Heather Clark, John Chandler, and Jim Barry (eds.), *Organisation and Identities: Text and Readings in Organisational Behaviour* (London: Thomson Learning), 132-44.
Miller, George A (2003), 'The Cognitive Revolution: a historical perspective', *Trends in Cognitive Science,* 7 (3), 141-44.
Moore, Michael S (1993), *Act and Crime: The Philosophy of Action and its Implications for Criminal Law* (Oxford: Oxford University Press).
Moors, Agnes (2013), 'Automaticity', in Daniel Reisberg (ed.), *The Oxford Handbook of Cognitive Psychology* (Paperback edn.; Oxford: Oxford University Press), 163-75.
Morewedge, Carey K and Kahneman, Daniel (2010), 'Associative processes in intuitive judgment', *Trends in Cognitive Sciences,* 14 (10), 435-40.
Morris, Katherine (2004), 'Introduction', in Katherine Morris (ed.), *Wittgenstein's Method: Neglected Aspects* (Oxford: Blackwell), 1-18.
--- (2007), 'Wittgenstein's Method: Ridding People of Philosophical Prejudices ', in Guy Kahane, Edward Kanterian, and Oskari Kuusela (eds.), *Wittgenstein and his Interpreters* (Oxford: Blackwell), 66-87.

Morse, Stephen J (1991), 'The "Guilty Mind:" Mens Rea', in Dorothy K Kagehiro and William S Laufer (eds.), *Handbook of Psychology and Law* (New York: Springer Science + Business Media), 205-29.
--- (1999), 'Craziness and Criminal Responsibility', *Behavioral Sciences and the Law,* 17, 147-64.
--- (2003), 'Inevitable Mens Rea', *University of Pensilvania Law School, Faculty Scholarship 526* (University of Pennsylvania).
--- (2007), 'The Non-Problem of Free Will in Forensic Psychiatry and Psychology ', *Behavioral Sciences and the Law,* 25, 203-20.
Mussweiler, Thomas, Englich, Birte, and Strack, Fritz (2004), 'Anchoring Effect', in Rüdiger F. Pohl (ed.), *Cognitive Illusions* (Hove and New York: Psychology Press), 183-200.
Myers, Greg (1989), 'The pragmatics of politeness in scientific articles', *Applied linguistics,* 10 (1), 1-35.
Neely, James H (1977), 'Semantic Priming and Retrieval from Lexical Memory: Roles of Inhibitionless Spreading Activation and Limited-Capacity Attention', *Journal of Experimental Psychology: General,* 105 (3), 226-54.
Neumann, Odmar (1984), 'Automatic Processing: a review of recent findings and a plea for an old theory', in W. Printz and A. Sanders (eds.), *Cognition and Motor Processes* (Berlin: Springer), 255-93.
Nickerson, Raymond S (2008), *Aspects of Rationality; Reflections on What it Means to Be Rational and Whether We Are* (New York/Hove: Psychology Press/Taylor and Francis).
Nicolson, Adam (2014), *The Mighty Dead: Why Homer Matters* (London: William Collins).
Nisbett, Richard E and Wilson, Timothy (1977), 'Telling More Than We Can Know: Verbal Reports on Mental Processes ', *Psychological Review,* 84 (3), 231-59.
Nisbett, Richard E, et al. (1973), 'Behavior as Seen by the Actor and as Seen by the Observer', *Journal of Personality and Social Psychology,* 27 (2), 154-64.
Nunan, David and Choi, Julie (eds.) (2010), *Language and Culture: Reflective Narratives and the Emergence of Identity* (New York: Routledge).
O' Callaghan, Casey (2012), 'Perception', in Keith Frankish and William M Ramsey (eds.), *The Cambridge Handbook of Cognitive Science* (Cambridge: Cambridge University Press), 73-91.
Oaksford, Mike and Chater, Nick (2010), 'Cognition and Conditionals: An Introduction', in Mike Oaksford and Nick Chater (eds.), *Cognition and Conditionals: Probability and Logic in Human Thinking* (Oxford: Oxford University Press), 3-36.
Oaksford, Mike, Chater, Nick, and Stewart, Neil (2012), 'Reasoning and Decision Making', in Keith Frankish and William Ramsey (eds.), *The Cambridge Handbook of Cognitive Science* (Cambridge: Cambridge University Press), 131-50.
Olson, David R (1994), *The World on Paper* (Cambridge: Cambridge University Press).
Pachur, Thorsten and Hertwig, Ralph (2011), 'On the Psychology of the Recognition Heuristic: Retrieval Primacy as a Key Determinant of its Use', in Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (eds.), *Heuristics: The Foundations of Adaptive Behaviour* (Oxford: Oxford University Press), 477-501.
Palahniuk, Chuck (2013), 'Nuts and Bolts: "Thought" Verbs.', 2015 (21/11/2015). <litreactor.com/essays/chuck-palahniuk/nuts-and-bolts-“thought”-verbs>.
Palgi, Yuval and Ben-Ezra, Menachem (2010), '"Back to the Future": Narrative Treatment for Post-Traumatic, Acute Stress Disorder in the Case of Paramedic Mr. G', *Pragmatic Case Studies in Psychotherapy,* 6 (1), 1-26.

Park, Heekyeong and Reder, Lynne M. (2004), 'Moses Illusion', in Rüdiger F. Pohl (ed.), *Cognitive Illusions* (Hove, New York: Psychology Press), 291.
Pennebaker, James W (1997), *Opening Up: The Healing Power of Expressing Emotions* (New York/London: Guilford Press).
--- (2004), *Writing to Heal: A Guided Journal for recovering from Trauma and Emotional Upheaval* (Oakland CA: New Harbinger Publications).
Peters, R. S. (1958), *The Concept of Motivation* (London: Routledge & Kegan Paul).
Phillips, Jean K Gilles and Woodman, Rebecca E (2008), 'The Insanity of the Mens Rea Model: Due Process and the Abolition of the Insanity Defence', *Pace Law Review,* 28 (3), 455-94.
Plous, Scott (1993), *The Psychology of Judgement and Decision Making*.
Polkinghorne, Donald E (1988), *Narrative Knowing and the Human Sciences* (Albany: State University of New York Press).
Popper, Karl (1992), *The Logic of Scientific Discovery* (London: Routledge).
Pritchard, Duncan (2004), 'The Epistemology of Testimony', *Philosophical Issues,* 14, 326-48.
Putnam, Hilary (1980), 'The Nature of Mental States', in Ned Block (ed.), *Readings in Philosophy of Psychology* (1; Cambridge MA: Harvard University Press), 223-31.
--- (1982), 'Three Kinds of Scientific Realism', *The Philosophical Quarterly,* 32 (128), 195-200.
--- (2002), 'The Nature of Mental States', in David J Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings* (Oxford: Oxford University Press), 73-79.
Ratcliffe, Matthew (2006), ''Folk psychology' is not folk psychology', *Phenomenology and the Cognitive Sciences,* 5, 31-52.
--- (2007), *Rethinking Commonsense Psychology* (Basingstoke: Palgrave Macmillan).
--- (2008), 'Farewell to Folk Psychology: A Response to Hutto', *International Journal of Philosophical Studies,* 16 (3), 445-51.
--- (2009), 'There are No Folk Psychological Narratives', in Daniel D Hutto (ed.), *Narrative and Folk Psychology* (Exeter: Imprint Academic), 379-406.
Ratcliffe, Matthew and Hutto, Daniel D. (2007), 'Introduction', in Matthew Ratcliffe and Daniel D. Hutto (eds.), *Folk Psychology Re-Assessed* (Dordrecht, NL: Springer), 1-22.
Recanati, François (1991), 'The Pragmatics of What is Said', in Steven Davis (ed.), *Pragmatics: A Reader* (Oxford: Oxford University Press), 97-120.
Reddy, Vasudevi and Morris, Paul (2009), 'Participants Don't Need Theories: Knowing Minds in Engagement', in Ivan Leudar and Alan Costall (eds.), *Against Theory of Mind* (Basingstoke: Palgrave Macmillan), 91-107.
Rennie, David L. (2012), 'Qualiative Research as Methodological Hermeneutics', *Psychological Methods,* 17 (3), 385-98.
Richardson, Frank C. and Fowers, Blaine J. (2010), 'Hermeneutics and Sociocultural Perspectives in Psychology', in Suzanne R. Kirschner and Jack Martin (eds.), *The Sociocultural Turn in Psychology: The Contextual Emergence of Mind and Self* (New York: Columbia University Press), 113-36.
Ricoeur, Paul (1984), *Time and Narrative*, trans. Kathleen McLaughlin and David Pellauer, 3 vols. (1; Chicago: University of Chicago Press).
--- (1991), 'Narrative Identity', *Philosophy Today,* 35 (1), 73-81.
Rieskamp, Jörg and Hoffrage, Ulrich (1999), 'When do People Use Simple Heuristics, and How Can We Tell', in Gerd Gigerenzer and Peter M. Todd (eds.), *Simple Heuristics that Make us Smart* (Oxford: Oxford University Press), 141-67.

Rieskamp, Jörg and Otto, Phillip E. (2011), 'SSL: A Theory of How People Learn to Select Strategies ', in Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (eds.), *Heuristics: The Foundations of Adaptive Behaviour* (Oxford: Oxford University Press), 244-66.

Robinson, John A and Hawpe, Linda (1986), 'Narrative Thinking as a Heuristic Process', in Theodore R Sarbin (ed.), *Narrative Psychology: The Storied Nature of Human Conduct* (New York: Praeger), 111-25.

Rorty, Richard (1970), 'In Defense of Eliminative Materialism', *The Review of Metaphysics,* 24 (1), 112-21.

Rosch, Eleanor and Mervis, Carolyn B (1975), 'Family Resemblances: Studies in the Internal Structure of Categories', *Cognitive Psychology,* 7, 573-605.

Ross, Lee (1977), 'The intuitive psychologist and his shortcomings: Distortions in the attribution process', *Advances in Experimental Social Psychology,* 10, 173-220.

Ross, Michael and Fletcher, Garth JO (1985), 'Attribution and social perception', *The handbook of social psychology,* 2, 73-114.

Ryan, Marie-Laure (2007), 'Toward a Definition of Narrative', in David Herman (ed.), *The Cambridge Companion to Narrative* (Cambridge: Cambridge University Press), 22-35.

Ryle, Gilbert (1949, 2000), *The Concept of Mind* (London: Penguin).

Salager-Meyer, Françoise (1994), 'Hedges and Textual Communicative Function in Medical English Written Discourse', *English for Specific Purposes,* 13 (2), 149-70.

Sarbin, Theodore R (1986), 'The Narrative as a Root Metaphor for Psychology', in Theodore R Sarbin (ed.), *Narrative Psychology: The Storied Nature of Human Conduct* (New York: Praeger), 3-21.

Sargent, T. J. (1993), *Bounded Rationality in Macroeconomics* (Oxford: Oxford University Press).

Saunders, Claire and Over, David E (2009), 'In Two Minds about Rationality', in Jonathan St B T Evans and Keith Frankish (eds.), *In Two Minds: Dual Processes and Beyond* (Oxford: Oxford University Press).

Saussure, Ferdinand de (1916/2011), *Coursein General Linguistics*, trans. Wade Baskin (New York/Chichester West Sussex: Columbia University Press).

Sayre, Francis Bowes (1932), 'Mens Rea', *Harvard Law Review,* 45 (6), 974-1026.

Scanlon, T M (2014), *Being Realistic About Reasons* (Oxford: Oxford University Press).

Schilling, Christopher J., Storm, Benjamin C., and Andersen, Michael C. (2014), 'Examining the Costs and Benifits of Inhibition in Memory Retrieval', *Cognition,* 133, 358-70.

Schlenker, Barry R. and Leary, Mark R. (1982), 'Social Anxiety and Self-Presentation: A Conceptualisation and Model', *Psychological Bulletin,* 92 (3), 641-69.

Schooler, Lael J. and Hertwig, Ralph (2011), 'How Forgetting Aids Heuristic Inference', in Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (eds.), *Heuristics: The Foundations of Adaptive Behaviour* (Oxford: Oxford University Press), 84-107.

Schröder, Hartmut and Zimmer, Dagmar (1997), 'Hedging Research in Pragmatics: A Bibliographical Research Guide to Hedging', in Raija Markkannen and Hartmut Schröder (eds.), *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts* (Berlin: De Gruyter), 249-70.

Schwarz, Norbert, et al. (1991), 'Ease of Retrieval as Information: Another look at the Availability Heuristic', *Journal of Personality and Social Psychology,* 61 (2), 195-202.

Seidenberg, M.S. and Tanenhaus, M.K. (1979), 'Orthographic Effects on Rhyme Monitoring', *Journal of experimental Psychology human learning and memory,* 5, 546-54.
Sellars, Wilfrid (1962), 'Philosophy and the Scientific Image of Man', in Robert Colodny (ed.), *Frontiers of Science and Philosophy* (Pittsburgh: University of Pittsburgh Press), 35-78.
Shapiro, Lawrence (2011), *Embodied Cognition* (Abingdon, New York: Routledge).
Shepherd, Joshua (2015), 'Conscious Control over Action', *Mind and Language,* 30 (3), 320-44.
Simon, Herbert A. (1955), 'A Behavioural Model of Rational Choice', *The Quarterly Journal of Economics,* 69 (1), 99-118.
--- (1991), 'Cognitive architectures and rational analysis: Comment', in Kurt VanLehn (ed.), *Architectures for Intelligence: The 22nd Carnegie Mellon Symposium on Cognition* (New York, London: Psychology Press), 25-39.
Skelton, John (1988), 'The care and maintenance of hedges', *ELT Journal,* 42 (1), 37-43.
Skinner, B F (1974), *About Behaviorism* (New York: Vintage).
Slovic, Paul, et al. (2002), 'The Affect Heuristic', *European Journal of Operational Research,* 177, 1333-52.
Smith, Adam (1776/1986), *The wealth of Nations (books I-III)* (London: Penguin Classics).
Smith, Eliot R and Mackie, Diane M (2007), *Social Psychology* (Hove, New York: Psychology Press).
Speekenbrink, Maarten and Shanks, David R (2013), 'Decision Making', in Daniel Reisberg (ed.), *Oxford Handbook of Cognitive Psychology* (Oxford: Oxford University Press), 682-703.
Spence, Donald P (1982), *Narrative Truth and Historical Truth* (New York: W W Norton).
Spencer, Cara (2007), 'Unconscious Vision and the Platitudes of Folk Psychology', *Philosophical Psychology,* 20 (3), 309-27.
Spiers, Hugo J., Maguire, Eleanor A., and Burgess, Neil (2001), 'Hippocampal Amnesia', *Neurocase,* 7, 357-82.
Stanovich, Keith E (2011), *Rationality and the Reflective Mind* (Oxford: Oxford University Press).
Steuber, Karsten R (2012), 'The Causal Autonomy of Reason Explanations and How Not th Worry about Causal Deviance', *Philosophy of the Social Sciences,* 43 (1), 24-45.
Stich, Stephen (1983), *From folk psychology to cognitive science: The case against belief* (Cambridge MA: MIT press).
Stich, Stephen and Nichols, Shaun (2003), 'Folk Psychology', in Stephen Stich and Ted A Warfield (eds.), *The Blackwell Guide to Philosophy of Mind* (Oxford: Blackwell), 235-55.
--- (2008), 'Excerpt from Folk Psychology: Simulation or Tacit Theory?', in William G Lycan and Jesse Prinz (eds.), *Mind and Cognition, an Anthology* (Oxford: Blackwell), 379-401.
Stiles, William B, et al. (2008), 'Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: replication in a larger sample', *Psychological Medicine,* 38, 677-88.
Sugarman, Jeff and Martin, Jack (2010), 'Agentive Hermeneutics', in Suzanne R. Kirschner and Jack Martin (eds.), *The Sociocultural Turn in Psychology: The Contextual Emergence of Mind and Self* (New York: Columbia University Press), 159-82.

Susswein, Noah and Racine, Timothy P. (2008), 'Sharing mental states: Causal and Definitional Issues', in Jordan Zlatev, et al. (eds.), *The Shared Mind: Perspectives on Intersubjectivity* (Amsterdam, Philadelphia PA: John Benjamins), 141-63.
Sykes, Elizabeth A. (2006), *The Psychology of Attention* (Hove and New York: Psychology Press).
Teigen, Karl Halvor (2004), 'Judgements by Representativeness', in Rüdiger F Pohl (ed.), *Cognitive Illusions* (Hove: Psychology Press), 166-82.
Terwee, Sybe J. S. (2012), *Hermeneutics in Psychology and Psychoanalysis* (Berlin: Springer Science & Business Media).
Thompson, Evan (2007), *Mind in Life: Biology, phenomenology, and the sciences of mind* (Cambridge MA/London: Belknap Press/Harvard University Press).
Thompson, Valerie A, Prowse Turner, Jamie A, and Pennycook, Gordon (2011), 'Intuition, Reason and Metacognition', *Cognitive Psychology,* 63, 107-40.
Thornton, Tim (2009), 'On the interface problem in philosophy and psychiatry', in Matthew R Broome and Lisa Bortolotti (eds.), *Psychiatry as Cognitive Neuroscience* (Oxford: Oxford University Press), 121-36.
Tulving, Endel (1972), 'Episodic and Semantic Memory', in Endel Tulving and Wayne Donaldson (eds.), *Organization of Memory* (New York: Academic Press), 381-403.
--- (2002), 'Episodic Memory from Mind to Brain', *Annual Review of Psychology,* 53, 1-25.
Tumulty, Maura (2012), 'Delusions and not-quite-beliefs', *Neuroethics,* 5 (1), 29-37.
Tversky, Amos and Kahneman, Daniel (1973), 'Availability: A Heuristic for judging Frequency and Probability', *Cognitive Psychology,* 5, 207-32.
--- (1974), 'Judgement Under Uncertainty: Heuristics and Biases', *Science,* 185, 1124-31.
--- (1983), 'Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement', *Psychological Review,* 90 (4), 293-315.
--- (1985), 'The Framing of Decisions and the Psychology of Choice', in George Wright (ed.), *Behavioral Decision Making* (Springer US), 25-41.
--- (1986), 'Rational choice and the framing of decisions', *The Journal of Business,* 59 (4), S251-S78.
Uleman, James S., Saribay, S. Adil, and Gonzalez, Celia M. (2007), 'Spontaneous Inferences, Implicit Impressions, and Implicit Theories', *Annual Review of Psychology,* 59, 329-60.
Van Boven, Leaf, Kamada, Akiko, and Gilovich, Thomas (1999), 'The Perceiver as Perceived: Everyday Intuitions About the Correspondence Bias', *Journal of Personality and Social Psychology,* 77 (6), 1188-99.
Varela, Francisco J, Thompson, Evan, and Rosch, Eleanor (1993), *The Embodied Mind* (Cambridge MA: MIT Press).
Verschueren, Niki and Schaeken, Walter (2010), 'A Multi-Layered Dual-Process Approach to Conditional Reasoning', *Cognition and Conditionals: Probability and Logic in Human Thinking* (Oxford: Oxford University Press), 355-70.
Vinden, Penelope G (1996), 'Junin Quechua Children's Understanding of Mind', *Child Development,* 67, 1707-16.
Von Eckardt, Barbara (2012), 'The Representational Theory of Mind', in Keith Frankish and William M Ramsey (eds.), *The Cambridge Handbook of Cognitive Science* (Cambridge: Cambridge University Press), 29-49.
Von Neumann, John and Morgenstern, Oskar (1947, 2007), *The Theory of Games and Economic Behavior* (Princeton NJ & Oxford: Princeton University Press).

Waismann, Friedrich (1968), 'How I See Philosophy', in Rom Harré (ed.), *How I See Philosophy* (London: Macmillan).
Waismeyer, Anna, Meltzoff, Andrew, and Gopnik, Alison (2014), 'Causal learning from probabilistic events in 24-month-olds: an action measure ', *Developmental Science*, 1-8.
Wason, Peter and Johnson-Laird, Philip (1970), 'A Conflict between Selecting and Evaluating Information in an Inferential Task', *British Journal of Psychology,* 61 (4), 509-15.
Watson, John B (1913), 'Psychology as the Behaviorist Views It', *Psychological Review,* 20, 158-77.
Wegner, Daniel M. (2005), 'Who is the Controller of Controlled Processes?', in Ran R. Hassin, James S. Uleman, and John A. Bargh (eds.), *The New Unconscious* (Oxford: Oxford University Press), 19-37.
Weinberg, Jonathan M., et al. (2012), 'Restrictionism and Reflection: Challenge Deflected, or Simply Redirected?', *The Monist,* 95 (2), 200-22.
White, P R R (2003), 'Beyond Modality and hedging: A dialogic view of the language of intersubjective stance', *Text,* 23 (2), 259-84.
Wilkes, Kathleen V. (1991), 'The Relationship Between Scientific Psychology and Common-Sense Psychology', *Synthese,* 89, 15-39.
Williams, Michael (2001), *Problems of Knowledge: a critical introduction to epistemology* (Oxford: Oxford University Press).
Williamson, Timothy (1996), 'Knowing and Asserting', *The Philosophical Review,* 105 (4), 489-523.
Wilson, Timothy D (2002), *Strangers to Ourselves: Discovering the Adaptive Unconscious* (Cambridge MA: Harvard University Press).
--- (2011), *Redirect: The Surprising Science of Psychological Change* (London: Allen Lane/Penguin).
Wittgenstein, Ludwig (1958), *The Blue and Brown Books* (Oxford: Blackwell).
--- (1975), *On Certainty*, trans. Denis Paul and G E M Anscombe (Oxford: Blackwell).
--- (1998), *Culture and Value* (Revised edn.; Oxford: Blackwell).
--- (2009), *Philosophical Investigations*, trans. G E M Anscombe, P M S Hacker, and Joachim Schulte (Fourth edn.; Oxford: Wiley-Blackwell).
--- (2012), *The Big Typescript TS 213*, trans. C Grant Luckhardt and Maximillian Aue (German-English Scholar's edn.; London: Wiley-Blackwell).
Wood, Frank, Ebert, Viola, and Kinsbourne, Marcel (2014), 'The Episodic-Semantic Memory Distinction in Memory and Amnesia: Clinical and Experimental Observations', in Laird S. Cermak (ed.), *Human Memory and Amnesia* (Hove, New York: Psychology Press), 167-93.
Yallop, David A (1971), *To Encourage The Others* (London: W H Allen).
Yeager, Daniel (2006), *J L Austin and The Law* (Lewisburg: Bucknell University Press).
Zadeh, Lofti A (1965), 'Fuzzy Sets', *Information and Control,* 8, 338-53.
Zamuner, Edoardo (2015), 'Emotions as Psychological Reactions', *Mind and Language,* 30 (1), 22-43.
Zimbardo, Philip G, Maslach, Christina, and Haney, Craig (2000), 'Reflections on the Stanford Prison Experiment', in Thomas Blass (ed.), *Obedience to Authority: Current Perspectives on the Milgram Paradigm* (Mahwah NJ: Lawrence Erlbaum), 193-237.
Zimmerman, Michael J (1997), 'A Plea for Accuses', *American Philosophical Quarterly,* 34 (2), 229-43.

Zlatev, Jordan, et al. (eds.) (2008), *The Shared Mind: Perspectives on Intersubjectivity* (Amsterdam, Philadelphia: John Benjamins).