Intentionality without intensionality

by
Matias Bulnes

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York
2013

This manuscript has been read and accepted for the

Graduate Faculty in Philosophy in satisfaction of the

dissertation requirement for the degree of Doctor of Philosophy.

Michael Levin

Date | Chair of the Examining Committee

Iakovos Vasiliou

Date | Executive Officer

John Greenwood

Stephen Neale

Rohit Parikh

Jesse Prinz

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

INTENTIONALITY WIHTOUT INTENSIONALITY

by

Matias Bulnes

Adviser: Professor Michael Levin

The main thesis of the dissertation is that the choice between individualism and externalism is a false dilemma. Individualism is the view that mental entities such as beliefs, desires, concepts, etc. are internal to an individual's body, individuated independently of what happens outside. Externalism is the view that they are external to the body, individuated by the relations between the individual and her environment. Both of these views disarm conundrums about the mind but are also plagued by their own problems. Yet the choice makes a material difference in nearly all mind-related problems. I do not wish to deny that they are mutually exclusive—I take this to be rather obvious. But in the dissertation I trace a route between them that avoids their perils.

I make one fundamental methodological assumption: mental entities are defined by their role in folk psychology. As a consequence, in adjudicating between competing hypotheses related to the nature of mental entities I rely on their power to explain and systematize folk psychology.

My strategy is simple. I begin with a diagnosis of the problems that plague individualism. I argue that by committing to the intensionality of the mental, individualism makes the content of mental entities (or what they represent) accessory to their individuation, contrary to folk

psychology. Then I weaken the individualist view just enough to avert this consequence without falling prey to the problems that plague externalism. The result is a view of mental entities as having individualist structure but externalist content. The objects immanent to the mental famously observed by Brentano emerge neither as neural symbols nor as external realities but rather as *virtual objects*, that is, as immanent to a characterization of brain states in terms of personal dispositions to interact with an environment.

ACKNOWLEDGMENTS

First and foremost, I would like to thank Michael Levin for his help both in substance and form with every chapter of this dissertation. Without his guidance this project would have had a very different fate. In fact, my debt to him goes beyond this dissertation: he has been a model for me as an academic. I would also like to thank Rohit Parikh for his detailed feedback throughout and especially for his technical support in Chapters 3 and 5. Thirdly, I would like to thank Stephen Neale for considerable assistance with Chapter 7. I also would like to thank John Greenwood for his encouragement throughout the doctoral program and help with the dissertation. Finally, I would also like to thank Jesse Prinz for his feedback and encouragement.

Table of Contents

Chapter 1

INTRODUCTION

Our vocabulary is imbued with mental words. Every day we talk about thoughts, beliefs, desires, intentions, fears, concepts, ideas, etc. Yet these notions do not fit squarely in a physicalist view of the world. In the past, logical positivism tried to dispense with these notions and thereby prove them inessential to a thorough understanding of humankind. But this attempt failed spectacularly. Eventually philosophers came to realize that somehow the mental is essential to our interaction with one another and view of the world. But this acknowledgement does not solve the problems that logical positivists sought to dissolve. Today these notions are widely accepted and used in scientific disciplines, based on the vague limits imposed by ordinary usage in combination with contextual standards of scientific respectability. For example, ordinary usage suggests that there are categories of mental objects, with concepts, ideas, etc. being components of beliefs, desires, thoughts, etc. Yet their true nature still eludes us.

But if we set out to unveil their true nature, where should we start looking? Traditional philosophical methodology suggests that the most reliable guide to the analysis of metaphysically ambiguous concepts is the role they occupy in our lives. On this point there is some consensus: mental notions exist to furnish psychological interpretation. Their fundamental role is to somehow inform our understanding and prediction of agents. David Lewis said as much in terms of linguistic meaning: the meaning of psychological terms must be determined by their role in the cluster of all (or nearly all) platitudes that compose folk psychology (1972). But this consensus has not materialized in a standard view of folk-psychological interpretation. Worse, even though philosophers of mind talk about psychological interpretation all the time, the absence of precise proposals about even its formal aspects is notorious. Yet we still stand a very small chance of hitting on the true nature of psychological entities, however meticulous our

speculations, as long as we are in the dark regarding psychological interpretation.

But surely Lewis' statement about the meaning of psychological terms is useless as a recipe to develop an account of psychological interpretation since it is impossible to catalog all the platitudes that compose folk psychology, let alone recognize patterns in them. How then can we construct a plausible account of folk-psychological interpretation without the very notions we ordinarily used to psychologically interpret each other?

First of all, we must get clear about the output of psychological interpretation. Probably because of the cultural prominence of the mind-body problem, philosophers have often assumed that psychological interpretation should work as a bridge between the two and so should issue in a hypothesis about the agent's bodily movements (e.g. Fodor 1968, Dretske 1988, Stich 1983, Kim 1984). But this goes against our every-day experience. Seldom when we know that an agent did something or we understand why she did it, do we thereby know how her body moved in doing so. In fact, oftentimes psychological interpretation enlightens us about actions whose physical aspects are manifestly irrelevant such as objecting to an argument or filing one's tax return.

The output of psychological interpretation usually is an event(s), bodily or otherwise, that some psychological states of the agent somehow explain (Belnap, Perloff & Xu 2001). When we understand an agent's action we understand why the agent brought about a certain event such as the request for a counterargument or the official incorporation in the IRS's system of her last year's tax-relevant information. Philosophers have called the subject matter of this understanding *practical reason*. But psychological interpretation also affords understanding of an agent's *theoretical reason*. When it does, it issues a belief, inclination, or other non-conative states. Thus, based on some of the agent's psychological states, we may understand her adoption of

other psychological states. However, because the role of mental notions in understanding theoretical reason seems irreducibly dependent on other mental notions, philosophers have focused their attention on the interpretation of practical reason in trying to provide a non-mental analysis of the mental notions. I will follow the tradition on this point.

When an agent has the intention to do something and also has, in the broadest sense, the ability to do it, it necessarily follows that she will do it. This logical connection between an intention to produce an effect, the ability to produce it and the occurrence of the effect is a pillar of the problem of evil. The differential between performing an action and failing to perform it is, therefore, the *ability* to perform it. When an agent intended to do something yet failed, it follows that she was, in some sense, unable. Sometimes ability may be a psychological matter, as when I intend to solve a chess problem but fail because I cannot get my *beliefs* about the position in order. But oftentimes it has nothing to do with psychological facts, as when I intend to kick a soccer ball from goal to goal in a soccer field but only manage to embarrass myself, or as when I intend to win the lottery but my number does not get picked. In cases like these there are no additional psychological facts that we need to know to understand why someone else actually kicked the ball from goal to goal or won the lottery. This suggests that there is a basic, unitary kind of understanding of an agent's practical reason which stops at the agent's adoption of the intention—whatever happens afterwards.

The natural target for the basic understanding of an agent's practical reason is the agent's deliberation leading to the adoption of her intention; call it practical deliberation. Plausibly, we understand why an agent adopted the intention to bring about a certain effect when we understand the deliberation that led her to adopting the intention. We have thus reduced the relevant parts of an account of psychological interpretation to an account of an individual

practical deliberation. This gives us a much needed hold on the challenge because deliberation is intimately connected with problem-solving, however it is realized in the brain. In fact, both practical and theoretical deliberations are arguably connected with solving optimality problems whose form we understand independently.

Decision theory provides us with a rubric for understanding practical deliberation.[1] In fact, the identification of the adoption of an intention to perform an action with reaching a decision what to do is natural on independent grounds. First, wherever deliberation is carried out successfully, it issues in a decision. This is not just true of deliberation in its psychological senses, but also in the legal sense of a jury's deliberation or any other. Second, though there arguably are intentions which are caused by non-psychological conditions and so are not identifiable with a decision, such as the intention to eat when we are starved or the intention to have sex when we are deprived, these are irrelevant to an account of deliberation because, not being identifiable with a decision, they are not the result of deliberation. In Hume's famous terminology, the intentions that matter for an account of deliberation are those that result from *reason* as opposed to *passion*. In fact, it seems to me that to say that in the latter cases *we adopt* the intention is slightly awkward since it is definitely not something *we did*. Still, the usage might be sustained by the social expectation that we control and take responsibility for these kinds of urges.

The application of decision theory to psychological interpretation has been hinted at many times before (e.g. Lewis 1983a, p. 114). In fact, some philosophers have already explored it (Schick 1991, Pettit 1991, Bermudez 2009). These are valuable efforts. But so much obscurity still remains that the result can hardly be considered solid grounds to settle other disputes within philosophy of mind. At any rate, those who have undertaken to explore the connection have not

[1] For theoretical deliberation we have Bayesian epistemology.

done so with avowed goal of illuminating other areas of philosophy of mind. But if my reasoning here is not mistaken, there may well not be another viable route to understanding psychological explanation than decision theory and, therefore, we should take the implications for a myriad of issues pending in mind-related disciplines seriously. For example, decision theory is eminently non-intensional because it has no place for representations.

In saying this I assume a particular construal of representations. I don't think that assigning a central role to decision theory in psychological interpretation is incompatible with our folk ontology of the mental as consisting of beliefs, desires, concepts, etc. As I will argue later (in chapter 5), my view is that ontological claims are harmless because they are trivial: anything can qualify as an entity (cf. Ladyman & Ross 2007, Ch. 4). Surely, beliefs, desires, thought, etc. are nominalizations of the verbs "to believe," "to desire," "to think," etc. and so of brain processes. But they are no less entities than conversations, purchases or falls. However, I distinguish between two types of representations: enactors and symbols. The first ones are representations whose individuation conditions are exhausted by their representational properties. The second ones are representations whose individuation conditions include more than just their representational properties (see chapter 4 below). If decision theory is central to psychological interpretation, beliefs, desires, etc. cannot be symbols: they cannot be individuated as such by anything more than their intentional content because decision theory deals with states of affairs, not with representations thereof. And if there cannot be two distinct mental representations (e.g. beliefs, desires, concepts, etc.) that have the same content, they cannot be intensional.

If decision theory proves essential to psychological interpretation, we will be forced to construe mental notions non-intensionally. This would upend the philosophy of mind which has

heavily relied on the intensional nature of intentional states since Frege and especially in the last decades under the dominance of computationalism. Frege explicitly introduced an extra determination for mental states over and above their content, namely, their sense. Computationalism altogether severs the individuation of mental states from their content: content turns out to be irrelevant to the role of mental representations in their brain (see chapter 4).

In this dissertation I set out to carry out the program I have summarized thus far. In chapter 2 I present a novel critique of individualism about mental states that exposes the crucial reliance of psychological explanation and the individuation of actions on the effects of an agent's deliberation. I do so by examining an intuitive property of folk-psychological explanations I call *self-sufficiency*. I argue that individualism cannot honor this property and work toward distilling an account of psychological explanation that does honor it, given some fairly standard assumptions. The account also hints at a distinction between psychological explanation and psychological interpretation, the latter concerned specifically with reasons for action. Later I will delimit the role of expected-utility maximization to interpretation.

Chapter 3 is a defense of the expected-utility conception of practical reason. I discuss the misconceptions that have blocked a wider adoption of this framework for understanding practical reason and argue that the best piece of evidence for it is the dearth of alternatives. I propose a new basic layout for its application that avoids all the misconceptions discussed. This layout, in turn, provides a straightforward systematization of the areas of folk psychology that crucially rely on the notion of rationality, given the assumption that folk psychology works by modeling the agent's deliberation. I conclude with applications of the proposal to four hallmarks of irrationality in folk psychology, among them akrasia and self-deception.

In chapter 4 I trace the fundamental problem with individualism to its commitment to

symbols. I define symbols as representations whose individuation is not exhausted by what they represent (or content). This commitment is of a piece with the assumption of the *intensionality* of the mental. In a nutshell, the problem is that the expected-utility framework has no room for symbols and so, given the argument of chapter 3, individualism is intrinsically incongruent with folk psychology. In contrast to symbols, an example of a representation fully individuated by what it represents is a statue of Napoleon. I call representations like this enactors and argue that folk-psychological entities such as beliefs and concepts are representations of this kind. Finally, I discuss the puzzles that have motivated the intensionality of the mental (and thereby individualism) from the expected-utility point of view and suggest that they can be accommodated with a simple move.

In chapter 5 I attempt to square the metaphysics that results from the view of agents as Bayesian optimizers suggested by the expected-utility framework. I argue that we should not dismiss an expected-utility account of practical reason because it does not fit in the standard armchair metaphysical picture dominating analytic philosophy in the last century. On the contrary, I argue that changing a simple yet fundamental metaphysical assumption about the concept of object used to describe the content of mental entities automatically dissolves some of the most vexing problems for the standard metaphysical picture. Moreover, such a change does not compromise our use of language and practices in the least. In fact, it resonates with some recent proposals in philosophy of science.

Using the conclusions of chapter 4, in chapter 6 I attempt to develop an account of psychological interpretation from the interpreter's vantage point. This construal of the problem of interpretation allows me to include merely possible objects in psychological interpretation in equal capacity as actual ones because it is indisputable that we can think and talk about non-

actual things. This, in turn, yields a novel account of how we manage to understand agents who suffer from ontological errors such as those who have motivated the assumption of the intensionality of the mental, brought to prominence by Frege. The idea reflects a theme that looms large throughout the dissertation: the traditional view of psychological interpretation in the model of garden-variety explanations—that is, as the observation of a phenomenon given independently of the interpreter—must be given up in favor of a view of psychological interpretation as aimed at coordination between interpreter and interpretee. According to this view, we cannot abstract from the interpreter in understanding what goes on in interpretation.

Finally, in chapter 7 I attempt to square the philosophy-of-language side of the equation. So far my concern has been entirely with psychological interpretation understood as an activity in which we engage independently of language. Yet something must be said about the communication of information valuable for psychological interpretation such as is accomplished in belief reports. Since all accounts of belief reports so far proposed rely, in some form or another, on the intensionality of the states reported, an alternative account is in order. Chapter 7 is my attempt to meet this challenge. I argue, contrary to conventional wisdom, that the long-standing puzzles associated with belief reports are not the result of their mysterious subject matter, namely, the mental. Rather, they arise from our failure to appreciate the wondrous flexibility of linguistic communication which research on the interface between semantics and pragmatics has made apparent in the last decade or so.

Chapter 2

INDIVIDUALISM AND THE METAPHYSICS OF ACTIONS

In this chapter I examine an intuitive property of folk-psychological explanations I call *self-sufficiency*. I argue that individualism cannot honor this property and work toward distilling an account of psychological explanation that does honor it, given some fairly standard assumptions. In doing so, my preference for an externalist individuation of intentional states will emerge unambiguously. The assumptions I rely on are fairly standard but not uncontroversial. Yet not always do I defend them from objections. My goal is an account of some areas of folk psychology consistent with our everyday practices rather than the deduction of an idealized psychology from first principles. The account will also have significant implications for our understanding of actions.

I. Self-Sufficiency

I will use as a model the following folk-psychological explanation: "Peter opened the fridge because he intended to get food." Let us not question the standard assumptions regarding explanations like this: 1) that it is an explanation (Churchland 1970); 2) that it is a *ceteris-paribus* explanation (Davidson 1970, Fodor 1974); 3) that the explanandum is an action, namely, Peter's opening of the fridge; 4) that the explanans comprises a state of Peter's (Davidson 1963), namely, the intention to get food; 5) that there is tacit information that supplements the explanans: at a minimum, that Peter believes the fridge to contain food.

Since I am interested primarily in the philosophies of mind and action, I will rely on a concept of explanation as weak and uncontroversial as possible. I will not assume that the explanandum can be deductively inferred from the explanans. Peter may have intended to get food and believed that there is food in the fridge, yet felt lazy and ordered Chinese. I will not

even assume that the explanans makes the explanandum likely. Peter might be the kind of person who never sets foot in the kitchen. I will simply assume that the conditional probability of the explanandum given the explanans is larger than the probability of the explanandum *simpliciter* (Salmon 1971, van Fraassen 1980). In short, the explanans need only raise the expectation that the explanandum will occur.

It takes no effort to appreciate the force of this explanation. You have a clear idea of why Peter opened the fridge even though I have given you no specific information about the situation. Peter might be a child or an adult, he might be blind or deaf, he might be depressed or euphoric, he might be at home or at his office, he might have been sleeping or sleep-deprived, and so on. The connection between the explanandum and the explanans appears to be independent of many particularities of the situation. With some exceptions: Peter must be a person, or at least an agent, with a fridge in his surroundings. One would naturally picture Peter as a normal man in modern times. But no particular information beyond this appears essential to the explanation. Put in terms of van Fraassen's famous account of explanation (1980), there obviously are many contexts in which our model can be a satisfactory answer to the question "why Peter opened the fridge?" and which do not include any background information about Peter's particular circumstances. Let us call this independence of the particularities of the situation the *self-sufficiency* of folk-psychological explanations.

In what follows, I will use this property of folk-psychological explanations against individualist accounts. After some preliminaries discussed in section 2, in sections 3 and 4 I lay down the basic challenge that self-sufficiency raises for individualism. In sections 5 and 6, I elaborate the view of mental states and actions resulting from the self-sufficiency of folk-

psychological explanations. Armed with this view, in sections 7, I trace the problem with individualism to an incompatibility with our practices and explanatory needs.

II. Effects of Actions

Davidson (1963) made popular the thesis that actions are individuated by their causes. The thesis that actions are also individuated by their effects, also defended by Davidson (1971), is less commonplace, but a good deal of evidence supports it.

First, the individuation of actions varies with context (Knobe 2006). On one occasion we may truly say that Peter and John were sentenced to life imprisonment because they did the same (or performed the same action), namely, committed premeditated murder. But on another occasion it may not ring odd to say that Peter and John did very different things, as one killed my friend while the other killed a stranger. The difference is explained by a contextual difference in the individuation of the intended effects of John and Peter's deliberations. Of course, the difference is also explained by a suitable contextual difference in the individuation of the very intentions John and Peter had. Yet there would hardly be any context in which we would say that John and Peter did the same thing on the grounds that both had the intention to kill Mary though only one succeeded.

Second, some actions do not even have an intention to individuate them. Immediately after I open my house door, I move my hand, flick the light switch, turn on the lights, illuminate the room, and alert the burglar that I am back (Davidson 1963). Our intuitions unmistakably sanction alerting the burglar as something *I did*, as opposed to something that *just happened*. Yet I had no intention or other mental state that accounts for the burglar's being alerted. This is what Feinberg (1972) has called the Accordion Effect. Of course, the Accordion Effect could be seen as no more than an anomaly. In fact, some have treated it like this (Cf. Bratman 2006). But it

would be a gain in parsimony to explain it as a function of the fact that effects weigh heavily in our commonsensical individuations of actions. Then perhaps it would not be entirely surprising that sometimes our judgments spill over to near, unintended effects of our intentional actions (Davidson 1971).

By contrast, effects are never missing in actions. That actions are usually exemplified with past-tensed sentences is no coincidence since the past-tense often conveys a particular change of course in the unfolding of the world. Unlike "Peter *plays* the piano," which refers to no particular action, "Peter *played* the piano" conveys that the piano was played in a particular occasion, and that it would not have been played had it not been for the intervention of an agent. This marks a difference with sentences that merely describe the current state of affairs such as "The piano is in the house." The change of course in the unfolding of the world embodies an effect and bespeaks a cause. True, action sentences always imply that some intention of the agent was a cause (Davidson 1971). But, as in the Accordion Effect, the intention not always psychologically explains the action's effect.

Since "Peter opened the fridge" doubtlessly stands for an action in our model explanation, the explanation must at least account for the effect of the fridge being opened. Let us call this effect the *minimal explanandum* of the explanation. Then individualism must be wrong, as I shall now argue.

III. Individualism and Self-sufficiency

By individualism I understand the thesis that mental states are individuated by properties intrinsic to the agent (Fodor 1981, McGinn 1982, Stich 1983, etc.). Most variants take the individuating properties of a mental state to be extrinsic to the state itself. Many believe them to be a brain state's power to cause, and its propensity to be caused by, other mental states. But

whatever they are, the properties of a mental state that determine what state it is must be independent of the agent's relation to any other objects, if individualism is true. Let us call properties of this kind *local properties*.

How would individualism account for our model explanation? The explanans would have to be events comprising only local properties of Peter's states. Yet the explanandum is clearly non-local, as it involves reference to the fridge. There seems to be only one possible way in which the expectation of even the minimal explanandum could be justified: the local relations invoked in the explanans provide the seed wherefrom to reconstruct the causal process leading to the explanandum. The fact that Peter intended to get food and believed that there was food in the fridge allows us to infer which mental state Peter will go into next. With some luck, this subsequent state will in turn be characterized in terms of its local properties so that we can infer what state Peter will go into next, and so on. Eventually, we will have reconstructed the entire causal process from Peter's initial state to his motor output. With the help of information relating mental states to bodily movement, we will have a fairly specific idea of how Peter's body will move. If the fridge happens to be in the right place at the right time in Peter's surrounding, we are in a position to infer that it will be opened.

But we face a problem: if the fridge is not in the right place at the right time the same sequence of bodily movements could lead Peter to drink from the toilet bowl. The output of the model explanation construed individualistically is a sequence of bodily movements whose effects in the world depend, among other things, on the geometry of Peter's environment. But since the explanation has to be self-sufficient, we should not need to know in advance which environment Peter is in. [1] We need more flexibility in our individualist explanation.

[1] Various versions of this objection have come up in the literature (Burge 1986, Peacocke 1993, Hornsby 1993, Wilson 1994). However, to the best of my knowledge, it has always been put more or less as a discrepancy between the narrow behavior

Individualism was inspired by the kind of computational functionalism advocated by Putnam 1963.[2] According to this view, a mental process is such that at every step the current mental state, in conjunction with the sensory input, issues the next state (and sometimes a motor output). The intervention of a sensory input at every step of the process provides the desired flexibility. Begin with Peter's intention to get food and his belief that there is food in the fridge. Add the sensory input that Peter is receiving at the time (for example, TV on in the living room) and infer the state he will go into. Repeat this sequence as needed and when a motor output is issued adjust Peter's spatial position and his sensory input accordingly. These calculations should lead us to the minimal explanandum that the fridge was opened whatever the geometry of Peter's environment.

I shall ignore the fact that this proposal is quite implausible as an account of what we grasp when we grasp the folk explanation. My interest lies in its self-sufficiency. On this score, the individualist explanation has come out of one hole just to fall into another, for the required flexibility necessitates the sensory input at every step along the way. Even if information about the sensory input did not require knowledge of Peter's particular situation, the mere information that Peter is having a sensory experience of a particular kind is obviously among the data that self-sufficiency is meant to exclude. In fact, you need no such information to grasp perfectly well why Peter opened the fridge.

Ultimately, self-sufficiency is a condition on the generality of folk-psychological explanations, and the problem with individualism is that the only avenue it leaves open for connecting the explanans and the explanandum is the reconstruction of the causal process that

---

that the individualist can get and the fact that folk-psychological explanation typically individuates behavior either intentionally or externally, or both. Putting the problem in terms of self-sufficiency has a couple of advantages. First, it dispenses with the obscure notion of behavior and replaces it with the more concrete event that serves as the minimal explanandum (for example, the fridge was opened). Secondly, it casts the problem at a more general level and presents this objection as a symptom of a deeper ailment. As a consequence, it hints at the heart of problem, as I shall try to show.

[2] Sometimes referred to as *machine functionalism*. See also Lewis 1972 and Armstrong 1968.

led from one to the other. Since this process will vary depending on the particularities of the situation, the explanation will vary accordingly. Individualism is therefore bound to violate the self-sufficiency of folk-psychological explanations.[3]

IV. Prediction and Accuracy

The generality bestowed on folk-psychological explanations by their self-sufficiency comes at a cost: folk-psychological explanations are usually frugal in what they tell us. Worse, their explanans have a fairly modest predictive power as they deliver at best a feeble expectation that the explanandum will occur. This is of little consequence when they occur in explanations, for in such cases all the relevant events are already known to have occurred.[4] In contrast, if all I know is that Peter has the intention to get food and believes that there is food in the fridge, I am justified in expecting him to go to the kitchen, open the fridge, etc. But consistent with all I know and with this expectation, he might also feel like having Chinese food and instead order in. That he will open the fridge need not even be likely.

Even so, it is unclear whether this should count as a disadvantage, for whatever predictive power folk psychology affords us, it places minimal demands on the information we need to feed it. This is an important lesson that the self-sufficiency of folk-psychological explanations teaches us.

[3] Throughout I avoid the debate over *narrow content* (for example, Fodor 1986). However, whether or not such a notion is coherent, it does not soothe the problems for individualism raised above. For one thing, Cartesian scenarios and Twin-Earth thought experiments show that one cannot in general read off external conditions from the narrow content of the agent's mental states. But more importantly, one can certainly not read off external conditions from the mere desire for food and the belief that there is food in the fridge, as self-sufficiency requires (Cf. Peter is just waking up from a nap and does not quite know in what room of his apartment he is).

[4] Cf. Railton 1978, van Fraassen 1980.

Individualists have generally failed to appreciate this point. Consider the following passage from Fodor 1981:[5]

> Suppose I know that John wants to meet the girl who lives next door, and suppose I know that this is true when 'wants to' is construed *opaquely*. Then, given even rough-and-ready generalizations about how people's behaviors are contingent upon their utilities, I can make some reasonable predictions (guesses) about what John is likely to do: . . . [He] is likely to call upon his neighbor. He is likely (at a minimum, and all things being equal) to exhibit next-door-directed behavior. (1981, p. 235; my emphasis)

Later in the same paper he adds:

> On the other hand, suppose that all I know is that John wants to meet the girl next door where 'wants to' is construed *transparently*; that is, all I know is that it's true *of* the girl next door that John wants to meet her. Then there is little or nothing that I can predict about how John is likely to proceed. (Ibid; my emphasis)

Here either Fodor misunderstands the implications of his individualist account or he is withholding information from us.[6] Fodor is arguing for a psychology that respects what he calls the *formality condition*, that is, a psychology that individuates mental states syntactically and, a fortiori, individualistically. Accordingly, in the passage, he uses the words "opaque" and "transparent" to refer to an individualist/syntactic individuation of mental states and to an externalist/semantic one, respectively.[7] He thus claims that it is the individualist individuation of mental states that sustains our *reasonable predictions*. But in light of the foregoing discussion, the individualist/syntactic individuation cannot ground the reasonable predictions Fodor claims it does—not without much more information about John's particular situation. Fodor uses a semantic characterization of John's mental state to illustrate the example and believes that he can

[5] See also Egan 1995.

[6] It is fair to say that Fodor has given up some of these views (1993).

[7] If there could be a doubt, he later on says: ". . . [It's] just as well that it is the fully *opaque* construal of mental states that we need since, patently, that's the only one that the formality condition permits. This is because the formality condition prohibits taxonomizing psychological states by reference to the semantic properties of mental representations and, at bottom, *transparency is a semantic (namely, nonformal; namely, non-syntactic) notion*." (p. 239; my emphasis).

turn the characterization into a syntactic/individualist one by simply construing the mental states *opaquely*. But even if such a proclamation somehow restricts substitution of co-referentials in descriptions of John's mental state, the mental state continues to be individuated in terms of, inter alia, the girl who lives next-door; else how on Earth Fodor would know even that there is a next-door to where John's behavior is directed! His predictions are therefore derived from a non-individualist characterization of John's mental state.

How would you present Fodor's case in strictly individualist (in fact, syntactic/computationalist) terms? Since you would have no access to *any* semantic interpretation of John's mental states, what you would have to produce is something like formulae using 0s and 1s, say 11110100, which somehow reveal John's current mental state's local causal powers in the computer language. Can you possibly infer from such formulae what John will do? Is there any event that you know is even likely to take place as a result of having this information? The only event you seem able to predict (provided you have access to the machine table) is that if John sees 01001011 he'll go into state 00101110 (and perhaps move his arm 10 inches up) and that if he hears 1010001 he'll go into state 1110100, etc. But since the information compatible with self-sufficiency does not include John's sensory input you cannot predict even this, let alone what happens beyond John's skin.

Some individualists have taken a more radical approach. Given determinism, there will always be an individualist explanation for every action. An individualist psychology, they say, would therefore render other explanations redundant. Whether it yields our folk-psychological explanations is irrelevant, for individualists believe that its accuracy makes individualism the only viable foundation for the development of a scientific psychology.[8]

[8] Versions of this argument can be found in Fodor 1968 (p. 42, in response to Peters 1958), Stich 1983 (Ch. 6) and McGinn 1982 (p. 208). For discussion see Pettit 1986.

To contradict such forecasts of scientific development would require their same audacity. Instead, suffice it to say that the foregoing discussion shows that this reasoning starts from a misevaluation of the merits of the individualist explanation vis-à-vis its competitors. For any extra accuracy of the individualist explanation comes at a far higher cost in information. So much so that science in its entirety has not been able to catch up to the individualist forecasts as we are yet to muster the information about the brain needed to produce one such explanation.

To get to the bottom of the problem that self-sufficiency reveals, in sections 5 and 6, I will attempt to shed some light on the limits that the self-sufficiency of folk-psychological explanations imposes on our notions of intention and action. In section 7, I will argue that these notions capture inherently non-local causal pattern and are therefore incompatible with individualism.

V. The Simple View

I will assume the Simple View, so called and rejected by Bratman (1984). This is the intuitive thesis that X'ing intentionally requires that one has the intention to X. I shall not attempt a full defense here (see McCann 1991) but I would like to emphasize the Simple View's simplicity. If true, we have a parsimonious unification of action and psychological explanation since "Peter X'ed intentionally" would both report an action and give a psychological explanation (namely, Peter X'ed *because* he had the intention to X). Moreover, given van Fraassen's account of explanation (1980), this unification sheds light on the conspicuous role of effects in individuating actions (section 2). According to van Fraassen, explanations are answers to why-questions which presuppose the occurrence of the effect and so, in our example, "Peter opened the fridge (intentionally)" is inseparable from the presupposition that the fridge was opened. If, on the other hand, the Simple View were false, we should ask: why do we so rarely use explanations like this;

for example, why do we hardly ever say that Peter opened the fridge because he had the intention to open the fridge? And the obvious answer, namely, that that is a roundabout way of saying that Peter opened the fridge intentionally, is not available. (Oftentimes the intentional nature of actions is only implicated, as in "Peter opened the fridge" in our model explanation. One must interpret this as an intentional action in order to make sense of the sentence in which it is embedded.)

The Simple View therefore has the consequence that reports of intentional actions are themselves psychological explanations and, hence, that the explanandum of our model explanation is a psychological explanation itself. The whole is thus a psychological explanation of another psychological explanation. To avoid an infinite regress let us stipulate that the explanation comprised in the report of an action is not an explanation of the action but merely of its associated effect, namely, the opening of the fridge. We may render the Simple View consistent as follows: Peter X'ed [A] intentionally iff (1) A was X'ed; and (2) this effect is *explained* by Peter's intention to A. (Here "S" stands for an agent, "X" for a verb and "A" for a term of some kind. I bracket "A" to indicate that this term is not always necessary.)

In honor of the Simple View, let us call sentences of the form "S X'ed [A] intentionally" *simple action sentences*. Let us agree that, like all sentences reporting effects, simple action sentences refer to an actual causal process culminating in the effect that A was X'ed. The difference between simple actions sentences and other sentences reporting effects is that simple action sentences cite the agent's intention as the initiator of the causal process and constrain the possible processes to an admissible range of alternatives.[9] It is thus that they explain the effect. Yet since the explanation works by referring to a causal process, we may also want to know why

[9] Here I mean to rule out cases of causal deviance (Davidson 1973, Mele 1992) but I shall not attempt to characterize the nature of the causal process characteristic of actions.

this process took place. Action explanations as our model are, therefore, explanations of the causal process leading from the agent's intention to the action's effect. That Peter had the intention to get food explains the occurrence of the causal process from Peter's intention to open the fridge to the opening of the fridge,[10] while the latter intention, in turn, causally explains the latter effect. This is the sense in which our model is a psychological explanation of another psychological explanation.

Simple action sentences are the simplest form of psychological explanation: they explain an effect by the presence of a psychological state. They say the minimum one can say about somebody's psychology to base an explanation on it. "Peter opened the fridge" says that the fridge was opened as the effect (in the right causal way) of Peter's intention to open it. But it does not say why Peter had this intention, what goal he sought with it or whether it was part of a larger plan. Folk-psychological explanations such as "Peter opened the fridge because he intended to get food" help us to broaden these simple explanations. They retain the explanation of the effect (the opening of the fridge) while expanding our understanding of Peter's psychology, of his motivations and goals (getting food), and so rationalize his action (Davidson 1963).

It is obvious that the explanation borne by a simple action sentence is self-sufficient in that the latter does not mention any particularities of the situation. Simple action sentences presuppose that some causal process within an admissible range of alternatives led from the agent's intention to the effect. But they are independent of exactly which causal process took place. Their truth-conditions only involve the agent's intention, the action's effect, and the

[10] This is congenial with van Fraassen's model of explanation as we can view the mention of Peter's intention to get food as answering the question why Peter opened the fridge (intentionally) when compared with the contrast-class, say, {staying in bed, watching TV, etc.}. However, if the contrast-class had been instead {ordering Chinese food, going to a restaurant, etc.} an answer to the same question would have required mentioning other aspects of Peter's psychological-make up beyond the scope of this article.

existence of some causal process leading from the former to the latter. I now want to go back to the individuation of the mental states cited in the explanans of folk-psychological explanations, only this time in relation to the explanations borne by simple action sentences.

VI. Intentions as Dispositions

What is the basic explanatory mechanism of folk psychology? How does the occurrence of an intention to X raise the expectation that X? If individualism is right, the intention to X *qua* mental state bears no special relation to the effect X. The explanation proceeds by a sequence of smaller steps leading to occurrence of X. This sequence is aided by much contingent information and so the relation between explanans and explanandum is synthetic. If this explanatory mechanism does not comport with the self-sufficiency of our folk-psychological explanations, the connection between them must be analytic (Fodor 1968, Ch. 1; Davidson 1963).

This possibility has been considered antagonistic to the view that explanans and explanandum are causally related (Melden 1961, Anscombe 1958, etc.). Hence, those who sympathize with an analytic connection between explanans and explanandum have generally denied that the agent's intention and the action's effect are causally related (Peters 1958, Ginet 1990, Wilson 1989, Sehon 1994, etc.). But I do not see how a causal relation can be denied here since I see no way an event could help to explain another, distinct event if they were causally unrelated (Lewis 1986a). Instead, I claim that the analyticity of folk-psychological explanations is compatible with the events being causally related.[11]

All that analyticity requires is that the explanans be individuated in such a way as to contain the information that the explanandum will occur (modulo qualifications). But it imposes no restriction on the kind of information that may be. This leaves one possibility open that makes

[11] The point is Davidson's (1963).

the explanation both analytic and causal: individuation of the explanans according to its power to cause the explanandum. The only alternative to individualism compatible with the fact that the explanans refers to a state of Peter's is that this is the state of being disposed to cause (in some special way) the opening of the fridge.

This thesis has been made unpopular by the absence of strict laws connecting agent's intentions and the effects of actions (Kim 1984, Davidson 1970). But while this requirement may have some plausibility for a scientific notion of causation, it obviously is out of touch with our commonsensical judgments regarding causes and effects. And since it is our commonsensical psychological explanations that I am interested in, I shall ignore it and leave the notion of causation unanalyzed.

For the purposes of illustration, however, a counterfactual analysis of causation along the lines of Lewis 1970 seems to me more commonsensical (Loewer & Lepore 1987). On this account, the event of Peter forming the intention to open the fridge caused the event of the fridge opening if and only if (a) the fridge would have been opened if Peter had the intention to open it; and (b) the fridge would not have been opened had Peter not had the intention to open it. (Of course, the first counterfactual is satisfied trivially when the two events are actual.)[12] If one prefers to take our counterfactual judgments as primitive, one can dispense with the notion of causation altogether and take conditions (a) and (b) as inherent in the explanation of why Peter opened the fridge.

Von Wright (1966, 1981) had already noted that action sentences generally involve a transition from one state to another. This insight has been developed into a logic of intentional

[12] These conditions embody a relation actually stronger than causation, one that Lewis calls *causal dependence*. Causation requires only that there be a chain of events between cause and effect all of whose links are causally dependent. Yet since typical folk-psychological explanations are self-sufficient and, therefore, our grasping them does not depend on our knowledge of the chain of events leading from explanans to explanandum, they must somehow rely on this stronger counterfactual relation, if they rely on any.

actions by a number of philosophers including von Wright himself (Aqvist 1974; Chellas 1969, Pörn 1974; Belnap, Perloff & Xu 2001). The basic idea is that an intentional action requires that the effect implicit in it counterfactually depends on the agent's forming the right intention. If, for example, the fridge was always open, or would have opened whether or not Peter formed the intention to open it, we would hardly say that Peter opened it (that is, that it was Peter's action). This idea is aptly captured by condition (b) above. Or if the fridge would always be closed despite Peter's intention, we would be left with no action, due to lack of an effect. Condition (a) implies that Peter's settling on the intention to open the fridge guaranteed that the fridge would be opened. True, causal deviant processes (Davidson 1973, Mele 1992) illustrate situations where an agent meets these conditions yet our intuitions clearly sanction her as not having performed an action. But taking something like the conjunction of (a) and (b) as necessary for intentional action has allowed these philosophers to shed light on an impressive number of areas of discourse—including imperatives and deontic sentences (Belnap, Perloff & Xu 2001).

Taking causation as primitive, however, suffices to accounts for the derivation of a mere expectation from the explanans of simple action explanations. That an entity has a disposition to cause X is compatible with X never occurring, since the manifestation of dispositions typically depends on background conditions. Yet, to the extent that we can expect these background conditions to hold, we can expect X to occur. Since the conditions on which the manifestation of intentions depends commonly hold, provided that we are not in possession of evidence that defeats this assumption, knowing that Peter has the intention to X justifies us in expecting that he will X.

These background conditions may include Peter's knowing how to X, his current ability to X, etc. (Churchland 1970) and are packed in the ceteris-paribus clause that any law of folk

psychology must have (Davidson 1970, Fodor 1974). If instead we positively know them to obtain, we can be certain that X will ensue.[13] This explains why our folk individuations of intentions so heavily rely on their content. If Peter and John have an intention such that the extant background conditions guarantees that they will X, then the intention they both have is the intention to X.

I do not claim this idea as my own (see, for example, Smith 1987, Papineau 1993) nor is it my goal to recast it. I simply want to work out the implications of the self-sufficiency of our everyday psychological explanations and this way of individuating mental states appears to be the only one capable of honoring it. In sections 7, I will argue that this self-sufficient, dispositional individuation of mental states is custom-made for the generalizations that matter to us.

VII. Simple Actions

Let us call the truth-maker of a simple action sentence a *simple action*. If what I have said so far is correct, a simple action is the manifestation of an agentive disposition. Since such a disposition is a disposition to cause the relevant effect in a special way, its manifestation is more than the mere occurrence of the effect. It requires that the state realizing the agent's disposition causes its effect through a characteristic kind of causal chain excluding causally deviant processes (Davidson 1973, Mele 1992). But the self-sufficiency of the psychological explanations comprised in simple action sentences guarantees that the exact causal chain of events leading from the state realizing the agent's intention to the effect is not part of their truth conditions. "Peter opened the fridge intentionally" would be equally true whether he stood up from his chair, walked to the fridge and pulled the door open with his right hand, or he helped

[13] Cf. the problem of evil.

himself to a stick to poke it open from his seat in the kitchen. These would be instances of the same simple action of Peter opening the fridge. Simple actions are therefore types, not tokens.

As with actions in general (Knobe 2006), the individuation of simple actions will vary with context. Oftentimes it will be natural to speak of a simple action as being independent of the agent. We may sometimes truly say that Peter and John performed the same simple action of opening the fridge. Sometimes we may even truly say that Peter and John performed the same simple action when they opened different fridges. This is due to the contextual sensitivity of the individuation of effects and intentions. When Peter and John open different fridges, they bring about effects that can be characterized in terms of each particular fridge. But also they bring about the effect of opening *a* fridge. Moreover, each one had the intention to open his particular fridge, but they also had the intention to open *a* fridge. This follows from viewing intentions as dispositions since dispositions are closed under logical entailment.

I believe that simple actions are often what we are talking about when we talk about intentional actions. The causal process bridging intention and effect seldom matters to our individuation. Individualism is, therefore, doomed to miss our generalizations over actions, as has been noted by, alas, one of its advocates:

> It simply will not do as an explanation of, say, why Mary came running out of the smoke-filled building, to say that that there was a certain sequence of expressions computed in her mind according to certain expression-transformation rules. However true that might be, it fails on a number of counts to provide an explanation of Mary's behavior. It does not show how or why this behavior is related to very similar behavior she would exhibit as a consequence of receiving a phone call in which she heard the utterance 'the building is on fire!', or as a consequence of her hearing the fire alarm or smelling smoke, or in fact following any event interpretable (given appropriate beliefs) as generally entailing that the building is on fire. The only way to both capture the important underlying generalizations ...and to see her behavior as being rationally related to certain conditions,

> is to take the bold but highly motivated step of interpreting the expressions in the theory as goals and beliefs (Pylyshyn 1980, p. 161).

Stich (1983), a staunch individualist, takes on Pylyshyn's challenge. He maintains that his syntactic theory of the mind does not miss any generalizations captured by our folk-psychological explanations. His strategy to defend this claim is to construct a parallel syntactic explanation for any folk-psychological explanation we can provide. Suppose we say that Mary heard the alarm and thus came to believe that the building was on fire; which, together with her long-standing desire to leave the building if it is on fire, led to her desire to leave the building which in turn caused her to leave the building; Stich will say that as a result of Mary's sensory state associated with the fire alarm she came to be in the B-state with syntactic form F, which together with her long-standing D-state with syntactic form F>L caused her to be in a D-state with form L which in turn caused her to leave the building. If instead Mary came to believe that the building was on fire as a result of smelling smoke, then Stich will issue an adjusted version of his explanation. He then concludes:

> What the various versions have in common, according to the [folk-psychological explanation] is that they all lead Mary to believe that the building is on fire, and this belief plays an essential role in the etiology of her fleeing behavior. On the purely syntactic explanation, there is a prima facie perfect parallel. What the various versions have in common is that *they all lead Mary to have the B-state F*, and this B-state plays an essential role in the etiology of her fleeing behavior. (Stich 1983, p. 176; my emphasis).

But Stich's parallel explanations are jointly incompatible with his syntactic individuation of mental states. Stich's alternative explanations require that *any* neural state in which Mary could have been right before getting the warning, in conjunction with the diversity of inputs that could have carried the warning, would have caused neural states, all of which belong to the same

syntactic type, namely, B-state F. Short of a miracle, the resulting states will differ significantly at the neural level since they are the effects of many very different afferent states. Because individualism requires that a B-state F is individuated by its local causal powers, it seems impossible for such a miscellaneous array of causes at the neural level to account for the fact that all the resulting states are instances of B-state F. Or, in other words, the semantic content carried by the inputs being unavailable to the individualist individuation, there is nothing to give unity to all the afferent states leading to the various instances of the B-state F. If individualism is to be compatible with Stich's explanations, B-state F must be individuated in terms of its effects.

However, the efferent processes of B-state F also vary dramatically depending on the sensory input Mary receives along the way. Mary might leave the building by running down the stairs or by jumping out of the window if the fire had reached the hallways. According to the syntactic taxonomy, the brain processes underlying these behaviors would most likely count as type-distinct because they would produce very different bodily movements, hence involve very different physiological responses, hence require very different causal powers. Had Mary chosen one or the other, the array of neural states she would have undergone would have been very miscellaneous indeed. Some of them might be closer to, say, playing basketball than to each other in terms of their local causal powers. Consequently, Stich cannot define the efferent process he calls "fleeing behavior" syntactically so as to bundle up all the miscellaneous afferent histories leading to the B-state F. B-state F and its successors in Stich's explanation can be individuated neither by their causes nor by their effects, as his syntactic theory would have it.

The problem is, as Pylyshyn's passage suggests, that we categorize reactions to danger by their intentional effects, not by the actual causal process bridging intention and effect, as witnesses the famous fight-or-flight response. This is unsurprising, since reactions like Mary's

are simply intentional actions prompted by some characteristic stimulus. Mary's fleeing behavior is actually the simple action associated with the intention to flee the building which reliably arises upon perception of threatening fire. This is why, as Stich notes, the belief that the building is on fire plays an essential role in the etiology of Mary's behavior. The effect (i.e. Mary running out of the smoked-filled building) is explained by the intention to leave the building alone, thus capturing our folk generalizations.

Such a coarse taxonomy of actions has its price, however. The simple actions of fleeing the building or opening the fridge lump together many miscellaneous causal processes. When I know that Mary fled the building, I still know little about what happened. Is this a consideration in favor of the individualist taxonomy? Quite the opposite; simple actions say little but this allows us to say much. In section 8, I will show how.

VIII. Complex Actions

Let us follow Bratman 1987 in viewing agents as planners and, accordingly, action as always associated with the execution of a plan. A plan can be viewed as a sequence of intentions each of which arising as the solution of a deliberation problem framed by the previous ones (Bratman 1987, Ch. 2). So the execution of a plan, to whatever point it succeeds, is the consecutive manifestation of each intention in the sequence into its effect. According to my characterization, each one of these manifestations, independently of the others, corresponds to a simple action. But the execution of the plan is not just a diverse collection of simple actions; each one of them is interlocked to its neighbors on both ends. On one end, the intentions are related by the agent's deliberation in the way just described. On the other end, their corresponding effects are related in inverse order: any given effect *ensures* the occurrence of the effect associated with the intention that preceded its own. Let us call any set consisting of two or more simple actions interlocked in

this way, a *complex action*.[14]

Like simple actions, complex actions are types. But they are more specific types and therefore subsume fewer tokens. As a result, complex actions carry more information about the actual causal processes that they subsume than simple actions and, hence, comprise better explanations of the final effect. Coarse though they are, simple actions do not exhaust folk-psychological explanation but merely are its building blocks.

Any execution of a plan, no matter how rudimentary, has numerous complex actions associated with it since any two simple actions in a plan can be interlocked. This is a consequence of the fact that any intention in a plan is a solution to a deliberation problem framed, recursively, by all the previous intentions. The "framing"-relation is therefore transitive. In one context, I can view Peter's intention to get food as framing the problem of where to go, but in another it might be more useful to view it as framing the problem of how to open the fridge. We may regard a complex action, A1, associated with a plan-execution as containing another complex action, A2, when all the simple actions that compose A2 are components of A1. But generally two complex actions need not even overlap on the simple actions that compose them to be part of the same plan-execution.

How fine-grained complex actions can be is a controversial subject, as it depends on the individuation of *basic actions* (for example, Baier 1971). Many muscle contractions appear to count as part of the execution of a plan, hence as simple actions (Danto 1965, Davidson 1971). But do neural events count as well (McCann 1974)? I do not have answers for these questions and will leave them open.

Still, suppose that we could produce a *total complex action* for a given plan execution,

[14] Things are actually more complicated, as some simple actions may be preliminary steps to the plan, there may be several simple actions which jointly (but not separately) ensure the manifestation of another one, or the plan may branch into two parallel sub-plans that later converge. For the sake of simplicity, I leave these cases out of the analysis but they can be incorporated without major disturbance of the fundamental idea.

that is, one that included all the basic actions as well as the non-basic ones. This would come close to specifying the actual causal process uniquely (I suppose it is conceivable that somebody could execute exactly the same plan to its nuances but this is surely unlikely). Even then the folk-psychological explanation arrived at would not be deductive in the sense that every step in the sequence occurs of logical necessity given the occurrence of its predecessors. The "framing"-relation is not a deductive one. For example, it does not follow logically from Peter's intention to get food that he will form the intention to open the fridge. The formation of a new intention also depends on the agent's beliefs in a way expressed by principles of rationality (Bratman 1987, Chs. 3 and 4). The good news is that the provision of information about the agent's beliefs, values, etc. can supplement the explanations comprised in complex actions by fleshing out the deliberation process framed by previous intentions wherefrom the new intention arose. I can say that Peter had the intention to get food, and his belief that he was broke, together with his belief that there always is food in the fridge, led him to form the intention to open the fridge, which together with his belief that the fridge is in the kitchen, led him to form the intention to walk to the kitchen, etc. Nor is this new explanation deductive. Whatever laws govern the deliberation process of an agent, they surely are *ceteris-paribus*, for there will always be further circumstances that could have prevented a certain intention from forming (Davidson 1970a, Fodor 1974). Even so, the more information one has about an agent and the deliberation process, the more one can narrow down the possible solutions, and so the more one can predict.

In sum, folk-psychological explanations are compoundable and supplementable with side information. This malleability sets them apart from other explanations such as the individualist one. Individualist explanations exhaust the causal chain underlying an action. No links can be added or subtracted from individualist explanations at the same functional level, for

if one link in the chain were missing the entire explanation would fail. By contrast, the self-sufficiency of folk-psychological explanations allows them to underdetermine the causal process in a way that can later be completed with more folk-psychological information. The speaker can thus choose the level of specificity of her account of a given action—subject, of course, to the limitations in the information she possesses.

IX. Back to Accuracy

Is folk psychology less accurate than individualist psychology? It depends on how the latter is conceived. Following Stich 1983, suppose that a syntactic psychology aspires to capture our folk-psychological processes of deliberation, only in syntactic terms. Surely then there will be a total complex action, as defined above, supplemented with beliefs and other information about the deliberation process, which breaks down this process into exactly the same mental state tokens as the envisaged individualist explanation would. This characterization of the actual phenomenon would be as fine-grained as the individualist one and, therefore, no less accurate. Any two intentional behaviors indistinguishable for this folk-psychological characterization would be indistinguishable for the envisaged individualist one and vice versa, despite their reliance on very different types.

Folk-psychological explanations may also be charged with inaccuracy stemming from their modest predictive power (section 4). This is a consequence of the dispositional character of these explanations. In general, dispositional explanations rely on ceteris-paribus clauses, as they depend on background conditions that need not be thoroughly specifiable. As a result, the occurrence of the antecedent justifies the expectation of the consequent but falls short of being sufficient for it. However, individualist explanations are also dispositional at every state in the chain leading from explanans to explanandum, and so they, too, rest on ceteris-paribus clauses

(Fodor 1974). The difference is a matter of degree: folk-psychological dispositions are more far-reaching than the individualist ones in that they characterize the agent's mental state in terms of more distant effects of the realizing state.

For example, more counterfactual circumstances can prevent the intention to open the fridge from manifesting than can block the manifestation of a disposition to bring about a certain computational state. If Peter intends to open the fridge, the fridge might still fail to be opened in a myriad of counterfactual scenarios, including one where Mary locked the fridge, one where Peter is restrained to his seat, one where he is blown away by a bomb right after forming the intention, etc. On the other hand, if Peter is in a given computational state and receives a certain sensory input, he might still not enter the following state if, for example, he has a heart attack or if he gets blown away by a bomb. But arguably all counterfactual scenarios where the computational disposition would fail to materialize also are scenarios where the folk-psychological disposition would fail. We may suppose that this grants the individualist explanation more predictive power.

However, a total folk-psychological description of the process realizing a given action would match the individualist explanation in predictive power. Suppose that we can produce such a description for Peter's intentional behavior up to the moment when he is about to open the fridge. Granted, the information required to produce such a description would be quite generous. Yet arguably it would be less than the information required to carry out the envisaged individualist explanation up to the same point. For instance, the folk-psychological explanation might very well dispense with information about the color of Peter's room, whereas there is no a priori reason to suppose that this information would be dispensable in the individualist explanation. To be sure, the computational process triggered by, say, the colors red and blue

might end in identical bodily movements in the kitchen, but there is no a priori reason to suppose that the computational processes leading there would be type-identical. Now if we were in possession of all the information required for a total folk-psychological characterization of the process right to the point when Peter is about to open the fridge, what degree of certainty could we have about the outcome? As much as the individualist could—right at that point. In fact, our folk-psychological characterization would grant us access to at least the intentional bodily movements Peter would be on his way to performing as part of his plan to open the fridge and this is the same information that the individualist would use in making her prediction (see section II).

No contest on the practical front. Information about an agent's beliefs and deliberative proclivities can increase the predictive power of folk-psychological explanations by narrowing down the possible solutions to her deliberation problems. This allows for the recycling of information. After forming a plausible hypothesis about the deliberation process Peter underwent today, save the relevant information about his beliefs and deliberative tendencies. Tomorrow when you have to deal with Peter again (or with someone relevantly similar), retrieve the information and you will be better at predicting him. Likewise, store such information as often as you run into Peter. Thus folk psychology provides us with the tools for profiling agents (by culture, personal beliefs, political views, aesthetic taste, etc.) and thus making our ability to predict an agent the better the more we exercise it. No such recycling is available to individualist psychology since every single piece of behavior, viewed through its lens, is *sui generis*.

I conclude with an application of the account to the moral evaluations of actions and to the teleological view of content.

X. Complex Actions in Moral Judgments

Consider the famous trolley problem:

1st Scenario:

A runaway trolley is about to run over and kill a group of five workers working unaware on the tracks. Peter finds himself in the control booth where he can divert the train to a side track. If he does, the train will run over another worker who is working on the side track. Would it be morally permissible for Peter to divert the train?

2nd Scenario:

A runaway trolley is headed to the group of workers but this time Peter is witnessing the situation from a bridge located between the current position of the train and the workers. Peter can save the five workers by pushing a big man who happens to be standing on the bridge onto the tracks, thereby derailing the trolley. Would it be morally permissible for Peter to push off the big man?

Strikingly, people systematically respond *yes* to the first question and *no* to the second. Yet the consequences of Peter's actions are identical in both scenarios. A plausible explanation of the difference is that people reason as follows. In the first scenario, Peter has to decide between two evils and does the right thing by choosing the lesser of them. In the second scenario, Peter has to decide between two evils too but choosing the lesser one would also involve violating other principles that in the first scenario do not come up—the no-battery principle, Kant's universalizability principle or what you will. But this requires that people distinguish between the structures of Peter's actions in the two scenarios. In particular, it requires that people view pushing the big man off the bridge as a component of what Peter did, in stark contrast to the relation he would bear to the casualty in the 1st Scenario.[15]

[15] The difference between pushing the man off the bridge and diverting the train to the side track is usually accounted for by appeal to the doctrine of double effect (for example, Mikhail 2007).

Mikhail 2007 and Knobe 2010 favor this solution and propose that when people face a moral problem they often base their judgments on structured actions. The idea is borrowed from Goldman 1970 who views structured actions as complexes of sub-actions related by some generative principle and represents them visually as action trees (Goldman 1970, Ch. 2; Knobe 2010). We usually verbalize structured actions with the help of expressions such as "by" or "in order to." We say, for example, that Peter saved the five workers *by* pushing another man onto the tracks; or that Peter opened the fridge *in order to* get food. These words are not exactly synonymous (Knobe 2010) but they are usually used interchangeably for the purposes of describing structured actions. You X by Y'ing when you Y in order to X; one is the converse of the other. But whatever the details of Goldman's account, the question is whether his structured actions correspond to complex actions as I have defined them. I maintain that they do.

First of all, Goldman's structured actions are composed by sub-actions which, however, are not themselves structured. To the extent that these are folk, unstructured actions they are what I have called simple actions. Second, the relation captured by "by" and its converse "in order to" strikingly resembles the relation I have conveyed by saying that simple actions *interlock*. Both are asymmetric, irreflexive and transitive (Mikhail 2007). These properties are possessed by the interlocking relation in virtue of the asymmetry, irreflexivity and transitivity of the "framing"-relation among the intentions characterizing simple actions of a plan (Bratman 1987). Moreover, "by" and "in order to" are expressions perfectly suited for describing plans. In explaining a recipe I might say "you will get the potatoes to have the right color *by* boiling them for 30 minutes" or "you must grind the corn *in order to* get the creamy consistence."

There is, however, one difference: Goldman's structured actions may include unintentional actions, while my complex actions are composed exclusively of simple actions

which are, by definition, intentional. But this is simply a difference of theoretical breadth. Goldman, Mikhail and Knobe are interested in actions *per se*, namely, in the overall role that they play in our conceptual scheme. I, on the other hand, am interested primarily in folk psychology and in actions only derivatively. Constraints on the concept of action relevant for their purposes, such as the way it interacts with attributions of moral responsibility, are irrelevant for mine. For instance, I think it rather obvious that the Accordion Effect is driven by our tendency to hold people responsible for more than they intend. Yet unintended effects of people's behavior are irrelevant to psychological explanation since, *ex hypothesi*, they lack such an explanation.

Even so, as Davidson (1971) argued, there is no unintentional action without, in the same instance, intentional action. This suggests that a full account of action must begin with an account of the relation between action and psychological explanation. In fact, it is quite plausible that the later computing stages of our moral judgments take as input the structure of intentional actions involved in a given situation (Mikhail 2007). Complex actions are, in sum, good candidates to serve as the backbone of structured actions—as a basic structure that is subsequently extended to incorporate unintended yet foreseeable effects in contextually relevant directions, as in the Accordion Effect.

XI. Reason Explanations

Following Anscombe 1958, suppose that reason explanations work by citing the *intention with which* the agent performed the action.[16] Then we face a problem. On the one hand, the fact that intentions are mental states makes it nearly inescapable that they must be causes of actions

[16] Of course, I am talking about internal reasons, that is, reasons the agent had in acting, as opposed to external reasons or objective reasons that the agent should have considered (for example, Raz 2005).

(Davidson 1963). On the other hand, our linguistic intuitions systematically sanction paraphrases of statements about the intention with which one acts into statements about one's goals in so acting (Wilson 1989, Ginet 1990, Sehon 1994). Are intentions causes or goals of actions? Given a counterfactual analysis of causation, they are both.

Reason explanations such as "Peter opened the fridge with the intention of getting food," are among the simplest compound folk-psychological explanations. Its explanandum is a simple action (that Peter opened the fridge—intentionally), and its explanans is simply an intention (to get food) different from the intention tacitly cited by the simple action (to open the fridge). Consequently, reason explanations lay down the agent's plan (or a self-contained portion thereof) in its barest form, adding the minimum to the fact that the simple action of opening the fridge is a component. In our model case, Peter's intention to get food frames a practical problem wherefrom his intention to open the fridge arises. Let us call these the framing intention and the framed intention, respectively.

Reason explanations tell us that the framed intention (to open the fridge) manifested while they remain silent as to whether the framing intention (to get food) did too. However, this silence does not undermine the account of reason explanation in terms of plans, for it is the simple action corresponding to the manifestation of the framed intention that is the explanandum, not the larger plan subordinate to the framing intention. The larger plan is a mere explanatory instrument. In fact, this assignment of roles marks reason explanations as teleological, for they explain the action in the explanandum, not by describing how its effect (the opening of the fridge) was brought about, but by describing what effect (getting food) the agent sought in performing the action. By contrast, consider a folk-psychological explanation such as "Peter got food by opening the fridge." The explanandum here is that Peter got food (intentionally) while

the explanans describes, if very roughly, how he brought about this action.

Teleological views of reason explanations have almost always been supported by the alleged impossibility of the framing intention serving both as a cause of the action and as its goal (for example, Ginet 1990). But proponents of teleological views have seldom had anything substantial to say about the metaphysical grounds on which the teleological explanation would work, if not causal grounds (Wilson 1989, Roth 1999). The structure of the agent's plan production provides such a ground. The framing intention is the simple action's goal in virtue of the characteristic way in which the framed intention arises in the agent's deliberation, namely, as a solution to the practical problem posed by the framing intention. This characteristic formation process renders the framed intention part of a larger plan, knowledge of which gives us a deeper insight into the agent's psychology by allowing us to move from a superficial intention to a deeper one in her means-end hierarchy—one that is, all else equal, more informative as to her characteristic desires, fundamental values, strengths and weaknesses, etc. By these psychological means reason explanations answer the question *why* the agent did what she did (Anscombe 1958).

This is why reasons are transitive. If to A is a reason why I B'ed and to B is reason why I C'ed, then to A is also a reason why I C'ed. Much as the intention to get food may be adduced as Peter's goal in opening the fridge, the intention to open the fridge may be adduced as Peter's goal in his walking to the kitchen. If so, then the intention to get food may also be adduced as a goal he sought in walking to the kitchen. Generally, given the transitivity of the "framing"-relation, any intention adduced as a goal of a given action is apt to be offered as a goal of any simple action following in the plan structure. Hence, reason explanations for a given action are in general not unique. Unless, of course, the simple action of interest is the first or second one in

the agent's entire deliberation, in which case it has one or no reason, respectively. However, many first intentions arising from biological drives such as hunger or sexual arousal still have a teleological explanation in terms of an evolutionary goal or the like. The psychology of agents is but a piece of a larger goal-directed whole.

I believe it clearest to reserve the word "goal" for agents. Agents have goals of all sorts; some of them psychological, some of them biological, perhaps some of them neither. But non-agents cannot have goals, though they can certainly have teleology. The function of the heart is to pump blood. But to my ears, that the heart's goal is to pump blood rings metaphoric. Likewise, intentions or actions lack goals, however much we sometimes demand to be told the point of a stupid action. I take this as convenient paraphrase of the question what goal *the agent had* in performing the stupid action. This paraphrase is also supported by the fact that usually in such cases we are talking about a simple action and, being a type, the same simple action can have scores of different goals on different occasions.

On the counterfactual analysis of causation favored here, the framing intention must be taken as a cause of the framed intention since the framed intention would not have been formed had the framing intention not been formed. But it is a special cause, distinct from, say, the way in which the Big Bang is a cause of the framed intention. The causal process leading from the framing intention to the framed intention must be amenable to being described as *a sequence of deliberation problems starting at* the framing intention. Perhaps there has to be a deliberation process describable in folk-psychological terms which leads from the framing intention to the framed one. Or perhaps the process must be fully realized in a neural network of some kind integrated into a bodily frame. Either way, no such description seems remotely plausible for the process leading from the Big Bang to the formation of the framed intention.

By the same analysis, the framing intention counts as a cause of the simple action that is the explanandum of a reason explanation. Simple actions were defined as the manifestation of an intention by some means within an admissible range of possibilities. The right manifestation of the framed intention is therefore the event that constitutes the explanandum of a reason explanation. Moreover, as a matter of conceptual necessity, the framed intention would not have manifested had it not been formed. And as a matter of metaphysical order, the framed intention would not have been formed had the framing intention not been instantiated. Consequently, as a matter of metaphysical order, the simple action would not have been performed had the framing intention not instantiated. In sum, reasons for actions comprise both their causes as well as their goals.

Chapter 3

FOLK PSYCHOLOGY, RATIONALITY AND DECISION THEORY

The connection between rationality and psychological interpretation has been emphasized by Davidson (1963) and Lewis (1974), among others. Though nowadays the connection is, nearly enough, conventional wisdom, it remains obscure. But there is a natural bridge for rationality and psychological interpretation: decision theory.

On the one hand, decision theory provides a framework for understanding psychological interpretation. We can understand an action by understanding the decision to perform it. For instance, I understand why Caesar crossed the Rubicon with his army by understanding the advantages and disadvantages he saw in doing so versus the advantages and disadvantages he saw in returning to Rome without his army. I do not know other ways of understanding an action. In particular, I do not know how to understand the bodily movements involved in Caesar's action because I do not even know which bodily movements he executed in performing it. But the understanding of the decision that led to the action surely is a psychological kind of understanding because it relies on psychological states of the performer such as her beliefs, desires, values, etc. Plus, such an understanding of the decision leading to an action seem to me exactly what I am looking for when I want to know why the agent performed it (Anscombe 1958).

On the other hand, decision theory is based on a notion of optimality which lends itself to identification with rationality. No doubt, rationality is a normative notion. When we accuse somebody of irrationality, we are thereby criticizing her. Because criticism normally carries the implication that, in some sense, one could have done better, it is natural to associate the concept of rationality to some sense of optimality. We can formulate the connection in sufficiently imprecise terms to preempt immediate rejection from any quarters:

R) An agent acts rationally iff the agent could not have done better.

Decision theory gives a more concrete meaning to this formula. First, the notion of acting involved in the left side corresponds to concrete actions as opposed to an agent's overall behavior. Second, the concept of the good involved in the right side corresponds to that of expected utility, which in turn breaks down into values and credence components. The good for an individual is the realization of what she values—where this not only depends on what she values but also on the chances of being realized. An agent acts rationally, therefore, when she could not have increased the likelihood of realizing more of what she values; in short, when she maximizes her expected utility.[1] Third, the counterfactual in the right side receives standard treatment but is restricted to worlds where the agent has the same psychological make-up. This follows from the fact that decision theory requires that the agent weighs all the options in the light of the same system of values and credences.

However, this unification of rationality and psychological explanation through decision theory is notorious for its lack of elaboration and advocates. Here I will try to correct this. First, I will cleanse the idea of various confusions associated with all or some the terms in the triad which eclipse its appeal. In the following section, I will show the productivity of the idea in resolving and clarifying a number of conundrums elicited by the intersection of these notions.

I. Decision Theory and Mental States

Some philosophers have taken decision theory to trade on beliefs and desires as the arguments of the value and credence functions. For example, in proposing their distinction between background and foreground theories of reasons, Pettit and Smith say:

[1] I ignore any distinction sometimes drawn between maximizing and optimizing (e.g. Mongin 2000) and assume that maximization does not require the existence of a unique maximum.

> As a foreground theory, decision theory …would counsel the agent to consider the state of his beliefs and desires with a view to determining in every choice the option which best serves those beliefs according to those desires, the option which maximizes expected utility. But it would be crazy to prescribe that an agent should deliberate only from considerations as to what he believes and desires, as distinct from considerations as to what is the case or what is desirable. We cannot seriously entertain the possibility that the rational agent should not consider the things he believes – that *p*, that if *p* then *q*, and so on – in deriving and justifying new beliefs, but should rather consider the fact that he believes that *p*, believes that if *p* then *q*, and the like. Neither can we countenance the possibility that he should not consider the factors that serve to justify and explain his desires – that an option will help a friend, that it will make him famous, or whatever – but should rather focus on the desires themselves. (Pettit & Smith 1997, p. 86)

But this passage confuses the subject matter of decision theory. Decision theory is not a theory of how to maximize one's utility in light of the mental states one has. It is a theory of how to maximize one's utility in light of how likely and valuable certain states of affairs are according to the evidence one possesses. Though subtle, the difference lies at the heart of the conflict between decision theory and the intensionality (with an "s") of the mental. If decision theory were, as Pettit and Smith suggest, a theory of how various mental states should give way to other mental states and, subsequently perhaps, to actions, the intensionality of the mental would perforce carry over to decision theory, given the standard folk individuation of mental states. As we saw, this is not the case.

Quite the opposite, decision theory trades entirely on states of affairs. What matters for the purpose of making decisions is not whether I believe that it is raining, but rather whether it is raining (e.g. Harman 1986). It is an epistemic karma all agents carry that the only way of assessing whether it is raining is by revising their subjective evidence—a karma immortalized by Descartes. But it is nonetheless the world, not ourselves, that constitutes the subject matter of decision making as construed by decision theory. It is, in fact, because decision theory is entirely

about how to reason one's way through the world that it is well-suited for psychological explanation. I think it quite plausible that the use of propositions instead of states of affairs in standard accounts of mental states is primarily a maneuver for marking sufficient epistemic distance between an agent's psychological processes and the world to allow for the representation of error. But if we go by the notions of probability and utility, state of affairs seem the natural match for decision theory (see above).

II. Lofty Values

A second confusion stems from the semantics of the word "desire" and the like. In its folk sense, "desire" usually refers to a longing or craving: an emotion associated with the possible occurrence of some state of affairs. However, the word has also acquired a technical meaning in the philosophy of mind and related disciplines, namely, that of a generic (and, perhaps, fundamental) pro attitude: the simplest, most basic motivational state (Davison 1963). Slote (1989) has argued against decision theory as a standard of rationality on the grounds that people do not always seek to satisfy their desires. But his argument equivocates because he inadvertently couches this claim in the first, narrower sense of "desire," while decision theory is most plausible when formulated in terms of the second, generic sense. Slote uses the following example:

> Imagine that it is midafternoon; you had a good lunch, and you are not now hungry; neither, on the other hand, are you sated. You would enjoy a candy bar or a Coca-Cola, if you had one, and there is in fact, right next to you, a refrigerator stocked with such snacks and provided gratis by the company for which you work. Realizing all this, do you, then, necessarily take and consume a snack? If you do not, is that necessarily because you are afraid to spoil your dinner, because you are on a diet, or because you are too busy? I think not. You may simply not feel the need for any such snack. You turn down a good thing, a sure enjoyment, because you are perfectly

> satisfied as you are. Most of us are often in situations of this sort, and many of us would often do the same thing. We are not boundless optimizers or maximizers, but are sometimes (more) modest in our desires and needs. But such modesty, such moderation, is arguably neither irrational nor unreasonable on our part. (Slote 1989, pp. 10-11).

In the generic sense, it is indeed perfectly rational for the agent to desire modesty and moderation and to behave accordingly—though we would not ordinarily (i.e. in the specific sense) call such a preference a *desire*. Rather, we would more naturally say that modesty and moderation are *values* that someone may have. This exposes the confusion from the point of view of decision theory. For the utility function is generic and so intended to include both longings and cravings, on the one hand, and lofty norms of conduct, on the other. In other words, it is perfectly consistent with decision theory that a person attaches much value to modesty and moderation. And this is all that a decision theorist would need to construe Slote's example as a case of utility-maximization.[2] In light of this, Slote's overall argument becomes a mere apology of the value of moderation.[3]

III. Boundless Maximization

But perhaps the most insidious problem regarding the connection between rationality and optimality results from obscurity on the exact way in which decision theory is to be applied to an agent's deliberation. Decision theory is not a theory of rationality, much less a theory of human deliberation. It is a theory of optimality in decision making with countless practical applications. Its application to the case of rationality may stand out because of its importance in other theoretical enterprises but should not be confused with the theory itself. As a consequence,

[2] Given Slote's insistence, in this example as elsewhere, that the agent is aware while deliberating that the snack is an option (Simon 1957). If the agent does not value moderation at all, nor is she sated, nor is she concerned about her figure or health, nor is he afraid of spoiling her dinner, it gets harder and harder to offer a rational justification for her turning down the snack.

[3] Slote also reformulates a couple of well-known conceptual conundrums for the advocates of rationality as utility-maximization but there also are well-known responses to them (Mintoff 1997).

decision theory can be applied to deliberation in more than one way, yielding more than one standard of rationality. Decision theory can be applied straightforwardly to an agent at any moment in time by taking the totality of her beliefs and desires at that moment (including degrees of confidence and desirability) and computing what action among the vast totality of options available would maximize her expected utility. And while this would surely set a standard for acting upon the world, it would also issue an unrealistic picture of agents' decision making since agents simply do not deliberate by considering all the options all the time. Alternatively, we could take into consideration the fact that agents face concrete practical problems and that their practical deliberation is shaped to solve these problems rather than to find the all-things-considered best use of their resources all the time. If so, we can apply decision theory to these concrete problems setting up a different standard of rationality.

In philosophy the debate over practical reason and rationality has been lively for a long time. Issues that have received considerable attention include the motivational force of reasons (Williams 1979, Smith 1987), the roles of beliefs and desires (e.g. Lewis 1988, 1996a) and conceptions of value that shape rationality (Parfit 1984). However, almost invariably these debates have proceeded on the implicit assumption that rationality calls for boundless, unrestricted maximization of the agent's expected utility. Consider, for example, Parfit's famous candidates for theories of rationality (e.g. Dancy 1997): the Self-interest theory (S), the Present-Aim theory (P) and Morality (M). Although they are couched in terms of beliefs and desires rather than the more technical credence and utility functions, all three conform to the optimality thesis as they all prescribe that one has most reason to do what maximizes *one's good*. They differ on what exactly *one's good* is. S construes it as one's total happiness during the course of one's life, P as the satisfaction of one's present desires, and M as the adherence to some

objective moral standards. Yet all three of them rely on the assumption that the maximization relevant to rationality is *total*, that is, over the domain of absolutely all of an agent's options.

In making this assumption the philosophical debate of the last five decades has followed the modern debate initiated by Hume. But it has been antithetical to nearly all other mind-related disciplines since Herbert Simon first published his ground-breaking work. Simon (1957, 1997) realized that the fact that actual agents are resource-bounded sets restrictions on the optimization processes they can undergo. Particularly, the combination of time and computational limitations determine that agents can never revise all the options available to them at a given time. Worse still, they usually cannot even gather the information that an option is available for all the options available nor, if they could, can they analyze and evaluate them. Simon concluded that it is a crucial aspect of our concept of rationality that we are resource-bounded and, therefore, that we must nearly always settle for what is sufficient and satisfactory over what is, in a boundless sense, best. In Simon's words, the goal of decision making usually is *satisficing* (a cross of "satisfy" and "suffice") rather than maximizing in the traditional sense.

The only plausible reason I can think of why philosophers have generally ignored Simon's work is that they have thought the boundless conception of rationality an ideal and, therefore, compatible with never being instantiated. If so—the argument would continue—Simon's conclusions do not pose a threat to the boundless conception for they trade on the imperfections which it is the role of the ideal to expose and which agents should, hopelessly, strive to abandon.

This reasoning is mistaken, however. While Simon's work does not by itself challenge the idea of optimality at the heart of the boundless view, it points at empirical facts about agents which seriously undermine its application to the understanding of rationality—even as an ideal.

In effect, if agents are resource-bounded it will not be rational for them to follow such an ideal willy-nilly since the very engagement in boundless maximization must at any given time be weighed against alternative uses of their limited computational resources that could be undertaken in the same time. Put bluntly, total maximizing may serve as a very abstract ideal reminding us of the rather obvious principle that the wider the domain of maximization the better the optimum, but it has little applicability as a standard of rationality for, in a rapidly changing world, striving to abide by it would too quickly turn the tables and become detrimental. For what would be most rational, according to the traditional boundless view, often changes faster than agents can compute and execute their decisions, leading to perennial vacillation and irresolution. Revising and changing all of one's decisions midway through execution is wasteful and, accordingly, perennial irresolution is a serious rational defect (Holton 1999). A perennially irresolute agent may never perform less than optimal actions but nor will she perform merely good actions and so it remains true that, in the long run, she could have done much better. By principle (R), boundless maximization cannot be rational.

IV. Plans and Agent Identity

In the case of bounded rationality, the exception does confirm the rule: Michael Bratman (1984, 1987, 1988) is the one philosopher who has developed the philosophical implications of Simon's ideas but his work has received more attention from computer scientists than from philosophers.[4] Bratman's work has vividly shown the importance of resource-boundedness not only in squaring the details of how agents make decisions in particular situations, but in shaping the very concept of agency. Specifically, he has shown how resource-boundedness informs an essential aspect of

[4] Judging from citations.

agency, namely, its consistency through time. Agents are not rampant utility-maximizers,[5] they are planners.[6] It is intrinsic to the agentive activity to device plans and follow them through. As Bratman emphasizes repeatedly, a plan is a commitment for the future, the locus of agents' projection and unity through time.[7] Without plans, agents would be a mere series of instantaneous utility-maximizers contiguous in time and space, however much each one valued the well-being of the others.

The idea that plans and optimality are intertwined in our concept of agency is borne out by recent experimental results. Children have been shown to recognize agency much before they have a full theory of mind (Wimmer & Perner 1983). In recent studies pre-theory-of-mind 1-year-olds have been shown to categorize non-anthropomorphic figures as agents when their behavior exhibits certain patterns of optimality (Gergely & Csibra 2003). More precisely, the experiments support the hypothesis that they see agency in objects that maximize efficiency in pursuit of a perceived goal (e.g. Pac-man). Crucially, the uncharacteristic appearance of the object suggests that it is irrelevant to agency categorization what the ultimate good is for them or whether their behavior maximizes their total utility. Rather, agency categorization in 1-year-olds seems to be driven entirely by the fact that the object's behavior fits into a putative plan, however sketchy and unjustified.

Parfit explicitly eschews plans in developing his preferred theory of rationality (the Present-Aim theory or P) when he excludes *derived* desires from those we have most reason to fulfill (1984, p. 117). Derived desires are desires that we have simply as means to achieve other

[5] Or, as Bratman calls them, *frictionless deliberators* (1987, p. 28).

[6] A feature that might very well turn out to be equivalent to being problem solvers, much emphasized by Simon (e.g. Newell & Simon 1972).

[7] In several parts, Bratman (1987) circumscribes his analysis to human agency specifically. But a conceptual leap from lower-order agency to human agency is implausible and so one must interpret this qualification as due to methodological frugality. To be sure, the plans that non-human animals are capable of are rudimentary compared to human plans (the dog plans to dig out the bone in order to eat), but they are plans nonetheless.

desires. According to Parfit, it falls short of the ideal of rationality to assign any special value to a goal I presently have in virtue of its being the path I previously chose to achieve another goal. So much so that what I have most reason to do altogether ignores these intermediate goals. Consequently, no commitment to a particular course of action is ever fully rational. Planning is futile from the point of view of Parfit's theory of rationality.

No wonder then that Parfit's preferred theory of rationality, *p,* portrays agents as a series of instantaneous selves. And he takes this to be a virtue of the theory as it allows him to accommodate the perceived bias toward the present, namely, that we can rationally care more about our present interest than about our future one. He claims that, given the similarity of indexicals "I" and "now," any grounds that justify a bias toward oneself (as opposed to others) will prove applicable to the case of the present self (as opposed to future ones). Since Parfit thinks the bias toward oneself is an indisputable fact (that is, that we do have more reason to seek our own well-being than we have to seek that of others) he concludes that the correct theory of rationality must recognize an admissible bias toward the present. This raises an insurmountable challenge for S since this theory, as defined by Parfit, is conceptually wedded to both the bias toward oneself and the time-neutrality of the individual good. In contrast, the Present-Aim theory embraces both biases, toward oneself and toward the present, and thus elegantly hurdles the obstacle.

But is hurdling this obstacle worth accepting the view that agents are mere temporal aggregations of spatially contiguous (underived-) desire-satisfiers with no metaphysical unity other than their contingent desires for each other's well-being? Not if you don't have to. And you don't: the planning conception of agency can have it both ways.

What makes an aggregation of temporally and spatially contiguous instantaneous selves a

single agent is the fact that their computational processings are interwoven: the decisions made by those selves occurring earlier in time structurally determine the decision problems faced by the ones following in the series (Bratman 1987, Bratman, Israel & Pollack 1988).This temporal knitting of computational processings accounts for both the spatial and temporal unity of agents. On the one hand, I am a different agent from you because my decisions do not structurally frame the decision problems you face and yours do not frame the decision problems I face. If I decide to get ice cream, it is me, not you, who later has to decide what flavor I will have.[8] On the other hand, I am the same agent now as I was an hour ago because some of the decisions I made then determine what decisions I am facing now. For instance, my previous decision to write this section determines that I am now facing some decisions as to what to say on this particular paragraph.[9]

Is it then rational to care about our present interest more than about our future one? Before answering this question, note that any bias toward the present should equally authorize one to prioritize the present interest over the future interest as over the past one. Parfit omits the past in his formulation of the problem probably because, on his theory, the past is settled and so irrelevant for the purposes of deciding what to do. However, on the view of agents as planners, the past does matter since, in carrying out plans, we are fulfilling goals that sprung in us in the past. Even when our interest in a certain goal has hitherto declined, if we are already embarked on a plan to fulfill it, the commitment implicit in the plan may suffice to push us through completion. This drive frequently manifests as a perceived obligation to finish our projects, even if we no longer give a damn about them.

So, is it rational to care more about our present interest than about our past and future

[8] However, this line might be blurred in the extremely rare cases of twins conjoined by the head. Recently, the New York Times reported a case of conjoined twins whose psychological identity is ambiguous ("Could Conjoined Twins Share a Mind?" published online on 05/25/2011)

[9] I do not propose this as an analysis of agentive identity through time but, at best, as a sufficient condition.

ones? No—if we are planners. Suppose that I believe that I will have a particular interest in the future. Barring cases of double lives or personality splits, this must be because I regard the future fulfillment of such an interest as a valuable goal now. I may not be embarked on the pursuit of the interest now, nor think it best for me to do so until certain conditions obtain. But unless I regard the fulfillment of the interest as valuable for a person under some conditions which I expect to meet in the future, my belief that I will have that interest is epistemically implausible.

The point is not always portrayed in ordinary parlance since the word "interest" is usually used to refer to the training of a skill. For example, a 20 years old person may want to have children in her 30s. Currently, we may say, she has no interest in children, but she believes that she will have an interest in about a decade. But this refers to the fact that the person is not currently seeking knowledge or proficiency in her handling of children. Strikingly, we would not say that she does not want children, but rather we would say that she does want them, only not now. In this case, "to want" conveys more accurately the idea of a goal or interest in Parfit's sense.

But if I do regard the future fulfillment of an interest as a valuable goal now, the rational thing to do is, intuitively, to craft a plan to fulfill it (provided it does not conflict with other goals that I regard as more valuable and/or viable). As Bratman observes, such a plan need not be precise in the least. It might simply consist in the intention not to adopt other plans that preempt the future fulfillment of the interest. Similarly, the commitment to a past interest, expressed in the current execution of a plan, normally determines the continuous entertainment of the interest in the present and future. In short, the view of agents as planners requires continuity and persistence of their interests. An agent whose interests are ever changing could not plan; hence, on this view, could not act, hence would not be an agent properly. This marks a fundamental

difference with Parfit's P, for while Parfit would probably agree that we cannot expect to have interests that we do not value now, his theory would not sanction an agent with ever changing interests an oxymoron.[10]

In sum, the question about an agent's future good does not come up for her for epistemic reasons: if she does not currently value a future goal, she simply does not know what she will want.[11] Hardly could it be rational to care about what we do not know. This partially explains the perceived bias toward the present.

But there also are more mundane factors that contribute to the perception. Most importantly, there are desires that appear to seize us in the sense of demanding our undivided attention to the point of ignoring much else that is important. Examples are the desires for food, warmth or sex that arise in cases of severe deprivation. The causes of the experience of being seized are most certainly neurological and, accordingly, the experience itself is largely irrelevant to the question of the rationality of giving these desires priority. Yet, at least sometimes, we deem it rational to give them priority over comparatively long-term goals. Plausibly, this is so when these desires are preconditions for the proper exercise of agentive skills and so when the neurological experience of being seized can be viewed as an evolutionary adaptation to the imperative need to fulfill them. For example, I may be starving to the point I cannot think clearly. Other things equal, it is rational in that situation to prioritize my interest in getting food—but not because I have it presently, nor because of the unpleasant experience of being seized, but rather because liberating myself from the neurological urgency and refueling my

[10] Although he might deny that such an agent would be a single person in any interesting sense of the word. Though subtle, it is crucial at this juncture to be clear about the distinction between theories of personal identity and the more fundamental theories of agency. See his views on personal identity in Parfit 1971.

[11] There is, however, the possibility that one knows by some sort of inductive inference that one will have an interest that one cannot presently conceived of herself as valuing. However, this case is relatively uninteresting for, ex hypothesi, the agent would have to presently repudiate her future self for acquiescing in such an interest. The TV show "This American Life" reports just such a case in a short episode about a proudly geeky teenager who self-imposes chastity for life on the (probably justified) grounds that sex makes people act stupidly. When asked to say something to his 30-year old self in case he yielded to temptation, he looks to the camera and says: "you should be ashamed of yourself!"

body are preliminary steps to the fulfillment of all my future interests, hence steps in any plan I may have. In this modest sense only it is perfectly rational to prioritize one's immediate interest over one's future interests or plans.

V. Decision Theory and the Structure of Thought

Losing sight of the fact that decision theory is a general theory of optimality in decision making rather than a theory of human deliberation may easily lead one astray into the boundless conception of rationality. For decision theory appeals to only two determinants of decisions, namely, the value and credence functions, which are naturally identified with beliefs and desires in our folk psychology. As a result, decision theory provides no psychological elements to build plans into the standard of rationality (Bratman 1987, Ch. 1, n.11). If plans reduce to beliefs and desires, there is nothing to give them a special place in deliberation over and above any other beliefs and desires that are not part of this or that plan. In other words, all beliefs and desires should be equal before the self: our deliberation should not favor a group of them, even if a plan supervenes on this group. Since viewing beliefs and desires as the ultimate propositional attitudes to which all others reduce has been the dominant trend in philosophy of mind in the last decades, plans have had their fate sealed in accounts of rationality.

Conversely, recognition of the characteristic role of plans in deliberation leads to a *reductio* of the view of the mind as, at bottom, a shapeless mass of beliefs and desires. Endowing this mass with structure takes only the addition of intentions as *sui generis* mental states.[12] Intentions are the building blocks of plans, states that comprise agents' endeavors (namely, their ends) wherefrom the concrete practical problems they face unfold. Having an intention to X typically brings out the problem how to X, and arguably there are no practical problems

[12] See Bratman's discussion of attempts to reduce intentions to beliefs and desires (1987, Ch. 1).

(perhaps, no problems at all) without an intention or goal of some kind. In what follows, I will assume that plans are sequences of decisions each issuing in an intention as the means to achieving the previous intention in the sequence. I will also assume that agents are essentially planners.

Once we recognize structure in our mental life we may avail ourselves of an intuitive understanding of rationalizing explanations to trace the role of decision theory in assessments of rationality. Davidson 1963 set out to provide just that intuitive understanding. He famously claimed that "[w]henever someone does something for a reason… [s]he can be characterized as (*a*) having some sort of pro attitude toward actions of a certain kind, and (*b*) believing… that his action is of that kind." (2006, p. 23). Davidson made this claim while elaborating on the idea that by recognizing the reason(s) why the person did what she did, we rationalize her action and, thus, come to understand it, as it were, from her point of view. Both conditions are too weak, however, as they are compatible with failures to rationalize (hence, understand) an action.

First, suppose that somebody performed an action A1 because she had a pro attitude toward actions of kind S (or the intention or goal to bring about S) and believed A1 to be of kind S. These facts are compatible with her also believing that there is another action, A2, which is also of kind S and which she thinks, all things considered, better than A1. In such a case, we still fail to understand why she performed action A1, as opposed to A2 or some other action. In other words, A1 rings irrational despite conditions (*a*) and (*b*) obtaining. What is missing in Davidson's condition (*b*) is the maximizing nature of the action-selection process. We understand why somebody performed an action (hence, rationalize the action) when we view the action not only as of the right kind, but as (one of) the best action(s), in the agent's eyes, of that kind that occurred to her.

Second, suppose that somebody performed an action A1 because she had a desire D1 to perform an action of kind S1 and believed that A1 was of kind S1. This is compatible with her also having had another, stronger desire D2 to perform an action of kind S2, believed that A2 is of kind S2 and, moreover, believed (perhaps, even known) that S1 and S2 are incompatible, that is, that the instantiation of one precludes the instantiation of the other. If we knew all this, we would still fail to understand why the person performed A1, as opposed to A2, for it is S2 that she knowingly wants the most. A1 still rings irrational. The problem in this case is that Davidson's condition (*a*) is too lax, his so-called pro attitudes letting in mental states that do not carry the slightest commitment to action (2006, p. 23). There is nothing exceptional about having incompatible desires, everybody has them all the time. Therefore, the mere fact that I desire something does not justify me in acting on it. In order to make Davidson's account work we need to allow only mental states which already presuppose a commitment to action (Bratman 1987) so that they screen off questions about the rationality of adopting the commitment in the first place. What we need, in short, is intentions or goals instead of Davidson's ecumenical pro attitudes, for having an intention or a goal presupposes that I have already adopted a commitment to act in its pursuit and, thereby, allows us to separate rational evaluations of my decisions in seeing it fulfilled from evaluations of my decision to adopt it.

With these two corrections, Davidson's account yields something like this: an agent performed action A rationally or, as we sometimes say, for a reason iff (a') she had an intention or a goal, and (b') she considered A best among the options she saw as means to fulfill that intention or goal. To use Davidson's own example, I flick the switch and thereby turn on the lights. The rationalization of my flicking the switch is that I intended to turn on the lights and believed that my flicking the switch is the best action at my disposal whose effect will be that the

lights go on (2006, p. 24). Here the optimality condition built into (b') seems superfluous and this is why the example suited Davidson's weaker account. But suppose there are two switches for the same light and so flicking either is an action of the kind that will bring about my goal. When several possible actions of the right kind suggest themselves, the possible worlds each action brings about will differ in scores of properties which I need not be indifferent to. For example, switch 1 may be located out of my reach on the opposite corner of the room while switch 2, though within my reach, may be in bad shape so that flicking it involves some risk of suffering an electric shock. Once I have recognized the options I have, I will evaluate them and choose the one I deem best. In addition to turning on the lights, I may also intend to get a beer from the fridge located on the opposite corner of the room by switch 1 and, accordingly, resolve that standing up and flicking switch 1 is the best action at my disposal.

In the language of decision theory, (a') translates into a construal of the optimality condition defining the standard of rationality for a decision as relativized to a previously adopted intention or goal. I will call it the intention or goal that dominates the decision. This is consonant with the view that agents are not rampant utility maximizers, that is, whose unique goal at every point in time is to maximize their overall expected utility. Rather, agents are understandable as maximizers of their expected utility given their commitment to a certain intention or goal. On the other hand, (b') translates into a construal of the domain of options relevant to the maximization of expected utility as restricted to those options viewed by the agent as conducive to the intention or goal dominating the decision. This is what Bratman calls the rational requirement of Means-End coherence (1987, Ch. 3). In sum, people are rationalizable to the extent that their decisions are solutions to concrete decision-theoretic problem framed by an intention or goal, the optimality condition applying to the selection of means dominated by the intention or goal.

We are now in a position to give a more specific interpretation of (R) stated in the previous section. The notion of the good remains expected utility as defined by the agent's beliefs and values. "The agent acts" remains interpreted as concrete, individual action rather than a set of them. However, the counterfactual in the right-hand side must be relativized not only to the agent's values and credences (and perhaps other cognitive features) but also to the agent's immediate intention or goal in performing the action.

VI. The Vagueness of "Rationally"

This may appear too narrow an understanding of rationality since clearly an action can be irrational simply because its immediate intention or goal itself is irrational even if the means selected to perform its immediate intention or goal could not have been bettered. I do not wish to deny that they can. However, I have been assuming that actions are the execution of plans which, in turn, are sequences of decisions each associated with an intention which dominates the subsequent decision. When a means is chosen for one intention or goal, the means becomes a subsequent intention or goal in the plan for which the agent will then select a subsequent means. Thus, the process repeats all the way down to basic actions, that is, actions that do not require decision-making (Danto 1965). But, crucially, this recursive structure entails that all the intentions and goals following any intention or goal in the plan are means to the fulfillment of the latter. This yields a justificatory structure according to which the adoption of any intention or goal is justified only if the adoption of every antecedent intentions or goals in the plan is.[13] My proposal is intended to apply recursively to the structure of plans, decision theory providing a mold that can be applied to all decisions in the sequence.

We must accordingly take the proposed decision-theoretic optimality condition as a bare

[13] Obviously, the reciprocate does not obtain because of condition (b').

minimum for evaluations of rationality in the sense that when we say, ordinarily, that we understand why somebody did something we imply, at least, that (a') and (b') obtain. But we can seek a deeper understanding of an action by applying the same optimality condition to previous steps of the plan of which the action is a part. In Davidson's example, the plan goes something like this: 1) My goal is, say, to read my book; as a means to achieve it, I choose to turn on the lights; 2) turning on the lights is now my goal; as a means to achieve it, I choose to flick switch 1; 3) flicking switch 1 is now my goal, etc. Rationalizing my flicking switch 1 requires, at a minimum, that you recognize (a') that my goal was to turn on the lights and (b') that flicking switch 1 was the best means I saw to achieve this goal. However, you might still, in a broader sense, fail to understand my action on the grounds that you do not understand why I had the goal to turn on the lights in the first place. Imagine, for the sake of the argument, that I am blind. If so, my goal of turning the lights on does not make sense as a means to read my book (we may suppose I use Braille) since it is means-end inconsistent with my goal and so should not have been in the domain of maximization. This simultaneously vitiates my action of flicking switch 1 despite its meeting the decision-theoretic optimality condition, for this, too, was a means to reading my book in the recursive structure of the plan and so means-end inconsistent with it.

This shows the need to expand our interpretation of (R). The proposed account teaches us that the possible worlds that enter the semantic evaluation of the counterfactual in the right side are determined by the explanatory interest of the interpreter. If you are seeking a minimal understanding of my flicking switch 1 you will be content with recognizing my goal of turning on the lights and flicking switch 1 as the best means to that goal. But if you are seeking a deeper understanding of my action you will want to scrutinize a larger portion of my plan in order to detect steps with respect to which *I could have done better*. Even so, this is cashed out in terms

of decision theory applied to the various decisions in the plan which enter into the evaluation. "Rationally" is thus vague.

In the extreme, someone may expect perfect rationality from an action. This means that she wants compliance with the optimality condition all the way up to fundamental goals or intentions, those that we have not as a result of deliberation. For example, someone may be satisfied with no less than knowing that my flicking switch 1 was the best means to achieve my intention to turn the lights on, which in turn was the best means to read my book, which in turn was the best means to fulfill my intention to relax, which in turn was the best means to fulfill my intention to feel good. At this point there is little point in demanding more, for no more can be given. Feeling good is not an intention we normally adopt to fulfill other intentions: it is, in Hume's sense, a passion. Consequently, the decision-theoretic optimality condition does not apply to it. Rationalization stops there.

But, as a matter of fact, our standards are seldom so high. Almost always we are satisfied with rather little. When we see unknown people walking in the street we feel perfectly satisfied with rationalizing their action in terms of their putative intention to go somewhere. And when they stop to ask for directions to a certain address we are perfectly satisfied with rationalizing their question in terms of their intention to go to the address, whatever they are planning to do there.

I do not wish to suggest that feeling good is an intrinsic goal or that it can only be sought for its own sake. It seems plausible to me that at least sometimes we take steps to feel good with no further goal in mind. But it also seems plausible to me that sometimes we take steps to feel good in order to achieve further goals (e.g. overcoming depression or resting from a rough week at work).

Nor do I claim that there are no other non-psychological senses of rationality that apply to fundamental goals. I take myself to have laid out a decision-theoretic framework to understand instrumental rationality. However, I do not rule out that there might be other non-instrumental sense of rationality which attach to certain intentions instead (or in addition to) the instrumental sense. For example, it may well be the case that other senses of optimality play a role in sanctioning intentions as rational or irrational in a non-instrumental way. Biological optimality (e.g. Friston 2010) could be one such sense. Or, perhaps, empathy plays a role in our non-instrumental judgments of rationality. Perhaps the intention to self-destroy is irrational in a non-instrumental sense because we simply can't relate to it. But whatever these non-instrumental senses are I claim they have nothing to do with psychological explanation. Their role must surely be to express expectations, demands or criticism, not to understand why the agent acted as she did.

VII. Some Applications of the Proposal

Higher-Order Rationality

Imagine that David is playing a chess game against his nemesis. As in tournaments, they are playing against the clock to make the game as fair as possible. Each one has exactly 30 minutes to make all her moves. Toward the end David has positional advantage but his time is running out. His nemesis, in contrast, has plenty of time left. The game comes down to a crucial point with only 30 seconds left on David's clock. If he makes the right move he can checkmate his nemesis in two extra moves. However, realizing that he will not have time to make all the calculations needed to determine the right move, David randomly chooses one of the two

contextually salient options.[14] As it happens, David loses the game. Afterward, you come up to him and tell him that you do not understand why he played the queen at that crucial point while moving the knight would have won the game in two moves. He explains that he sees that fact now that he is under no time pressure, but at that moment, he very much doubted that he would be able to arrive at the right assessment of his options before his time ran out and, accordingly, decided to take a rough estimate giving each equal chance of being the right move. He figured this would at least give him about 50% chance of winning. Now you understand why he did what he did, even though, relative to his values, beliefs, goals, etc. he did not choose the best option at his disposal and so his action did not comply with the decision-theoretic optimality condition.

Isn't this a counterexample to the decision-theoretic account of rationality I am proposing? I say it is not. Rather, the case illustrates the complexity involved in the psychological assessment of agents with higher-order thought. The key lies in David's belief that he would not have enough time to carry out the thought process necessary to determine which of two options leads to the winning position. This issues in a corresponding decision to ignore his knowledge and try the roughest of approximations, namely, give each option 50% chance. In other words, this is the case of a higher-order thought process (namely, a thought process about one's own thought process) co-opting the standard lower-order thought process, resulting in a lower-order decision that violates the optimality condition imposed by decision theory on the lower-order decision. I claim that the influence of higher-order thought processes can have this effect but the resulting decision will be rational if the higher-order decision itself complies with the decision-theoretic optimality condition. In such cases it remains true that the agent could not have done better.

[14] I have chosen rapid chess (i.e. 30 minutes per side) to illustrate the example for expository reasons. But this situation is, in fact, almost inevitable in blitz chess (i.e. 3 minutes per side).

The modulating effect that higher-order thought has in plan construction has long been recognized by psychologists. For example, research in the psychology of decision making has shown that, in adherence to the principle that a wider domain guarantees a better optimum, people tend to spend proportionally more time generating options than ordering and choosing among them, the more time they have to make a decision and, furthermore, the more importance they attach to it (Hogarth 1980). Therefore, it has been known for a while that humans have volitional control over procedural aspects of their decision making such as the generation of options that set up their domain of maximization. If so, even while finding fault neither with my choice of action to pursue a goal nor with the goal itself, you may still think the overall thought process deficient on the basis of my choice of time and effort devoted to generating options. To use our previous example, you may understand why I flicked switch 1 in the sense of understanding that this action was the best among those I considered. But you may still take issue with my having forgotten about the clapper switch because you think that it was sufficiently salient in the context or because you think that the decision mattered enough to deserve more time and attention than I gave it.[15] This is the exact opposite of the chess case: an action is found to be deficient even though it complied with the optimality condition on account of the higher-order decisions that shaped it.

Daniel Kahneman (2011) has brought this division of labor within the mind into sharper focus. For years Kahneman has been studying two decision-making systems in the human brain—System 1 and System 2, as he calls them. Scores of functional differences testify to the robustness of Kanheman's distinction. System 1 is automatic, ongoing, subconscious, parallel, associative, effortless, the default operator and commonsensically viewed as embodying

[15] I acknowledge that Davidson's switch-flicking example does not suggest any importance. However, imagine that I just had surgery on my leg and the doctor has insisted on not using it…

intuition. System 2 is deliberate, selective, conscious, serial, inferential, effortful, can supervise System 1 and is commonsensically viewed as embodying the self. Moreover, System 2 is believed to be younger than System 1 in evolutionary history, which sits well with its association with higher-order though—as this is a sophisticated ability present only in a few rather advanced animals, as witness the so-called mirror test (Gallup 1970). I am not suggesting a straightforward identification of System 2 and second-order thought, however. In fact, it seems that System 2 can enact several (perhaps all) intentional levels recognizing and shaping features of its own workings. The point is rather that much evidence shows that System 2 has the ability to co-opt and manipulate lower-order decision making. This offers a natural way of understanding what is going on in the cases that interest me here.

In fact, playing chess has been highlighted as an example of System 1 at work. Chess masters have an uncanny ability to recognize chess positions intuitively, that is, without a conscious, deliberate analysis. Of course, they can combine such ability with conscious analysis and they surely do so in tournaments. But they can bring System 2 to bear on their thought processes at various intentional levels. In particular, they can keep track of the time in speed chess and impose this restriction on all lower-order decisions including their intuitive analysis embodied in System 1. Though David is no chess master, plausibly, this is what is happening in his case.

As for the case of the clapper switch, recognizing the interaction between these two systems probably creates expectations in us as to how they are to be used according to the importance of various decisions. Coarse though it is, System 1 successfully handles most quotidian tasks. Opening doors, sitting down, standing up, grabbing objects, recognizing speech, producing speech, and even recognizing moods in others, all are normally dealt with masterfully

by System 1. However, System 1 is tuned to statistical success and occasionally leads us stray. When a decision comes up for which we should not take the chance that System 1 will fail, we are supposed to bring System 2 to bear on the process. If my decision to turn on the light had been one such, I should have brought System 2 to the task and am vulnerable to criticism for not doing it.

But aside from the illustration afforded by Kahneman's two-system model, I think it is quite plausible that the function of higher-order thought is to control lower-order thought. By "control" I mean the sense involved in *process control* in mathematical engineering, namely, the idea of a system keeping another system in check or, more technically, keeping its output within a given range of parameters. In this sense, the controlling system must process information about the controlled system and start from a predetermined measure of adequacy for the relevant output. No wonder then higher-order thought is higher-order, that is, having as its content the lower-order processes. On the other hand, the output which would be under control would be expected utility. Moreover, this function should certainly give higher-order decision the power to co-opt lower-order processing, for without such a prerogative the control system would be pointless. The fact that higher-order thought has the last word goes a long way to explain the commonsensical view that Kahneman's System 2 embodies the self. Also, this answers the question why David's consideration regarding time can co-opt his decision making in the chess game.

But most important for our purposes is the question about rationality: Why do we not view these cases as breaches of rationality—even despite the failure of the optimality condition at the lower-lever? The answer is that the whole point of process control is to further optimize performance. In the case of higher-order thought, if its function is to control lower-order

processes, it will generally better the output of the lower-order decision making process since the option of not interfering with the lower-order process is always available. However, what matters for assessments of rationality is not the impact of the higher-order process on the lower-order one but the performance of the integrated system.

Under certain conditions, it may be detrimental to even activate the higher-order, control processes. These conditions are given by the balance between the cost of time and resources resulting from higher-order control versus the gain in utility expected from exercising control. This implies rather intricate criteria. If, for example, time barely allows for the running of lower-order processes, activating higher-order processes may delay the output and force a very hasty decision. I have found myself in these situations when taking standardized tests. To be successful in these tests one must literally train the problem-solving skills targeted by a particular test, for one must be ready to intuit the answers with minimal control of what one is doing. One other hand, when time is even scarcer to the point it does not even allow for lower-order processes to run, a hasty decision may be better than no decision and higher-order processes should interfere blocking lower-order processing. This seems to be the case of David's chess game. Because the gain in expected utility is usually unknown a priori it may be hard to know when it is worth controlling (cf. Bratman & Israel 1988).

The proposed role for higher-order thought comports with the general account of rationalizations I am defending, since the performance of the integrated system of higher- and lower-order processing is measured in terms of expected utility. In other words, what makes an action rationalizable still is that all the decisions involved in it (whether higher- or lower-order) respect the decision-theoretic optimality condition. Yet overall evaluations are intricate, for they must take into consideration the levels involved in the structure of thought.

I have been assuming that higher-order thought processes differ from lower-order ones only in their content. In particular, higher-order thought processes should have the same plan structure that I have attributed to lower-order ones. The ultimate goal or intention of those plans, however, is always to aid the lower-order processes so as to optimize overall performance. But because they have the same structures they are evaluated in the same way. Decision theory provides a mold that we can use to assess the various stages of the plan. Any violation of the decision-theoretic optimality condition (relativized, of course, to the agent's credence and value system) by any decision in the higher-order plan structure translates into a breach of rationality. However, the overall assessment of rationality must also incorporate the function of the higher-order thought process in shaping the lower-order ones. Consequently, the lower-order processes are subject to the decision-theoretic optimality condition *unless* the condition is overridden by higher-order decisions. The result still is rational, provided that the higher-order decision to override the optimality condition at the lower-level itself conforms to the decision-theoretic optimality condition. This is what happens in David's chess case. David acted rationally, therefore, for it remains true that he could not have done better.

Inconsistency

Having contradictory beliefs is a flagrant form of irrationality because everything follows from a contradiction, and so a contradictory agent is in a position to infer anything, true or false. Things are worse if we assume logical omniscience, that is, the view that an agent believes all the logical consequences of her beliefs. Under logical omniscience, having merely inconsistent beliefs, namely, beliefs which together (strictly) imply a contradiction, is nothing short of believing the contradiction itself. On this view, all epistemic inconsistencies mask contradictory beliefs; hence they all are equally irrational. This is a reductio of logical omniscience. Inconsistencies seem to

be fairly ordinary and they do not seem to pose as big a threat to our overall psychological understanding of an agent as contradictions.

Unlike contradictory beliefs, inconsistent beliefs can come in bunches of more than two. For example, the belief that this or that grain of wheat is not a heap, the belief that this other grain of wheat and the previous one are not heap, the belief that this other grain of wheat and the previous two are not a heap, etc. together with the belief that all the grains together are a heap are jointly inconsistent, without any pair (in fact, any proper subset) of them being inconsistent. Worse, the addition of any belief to an inconsistent set of beliefs yields another inconsistent set of beliefs. However, I will call a set like this *weakly inconsistent* because it has a subset that is inconsistent. When talking about inconsistent beliefs simpliciter, I will therefore be talking about a minimal set of inconsistent beliefs.

Inconsistent beliefs quite directly compromise theoretical rationality, for they provide grounds to derive opposite beliefs. In terms of Bayesian epistemology, if $p_1, p_2, \ldots, p_n$ are inconsistent propositions, the conditional probability of one given the others is zero. Hence, evidently an agent cannot believe one by conditionalizing (correctly) over her total evidence, including the others. As a consequence, no subjective probability function can be defined for an agent who believes all of them. Such a function would assign the necessary proposition $\sim p_1 \vee \sim p_2 \vee \ldots \vee \sim p_n$ a probability smaller than 1 and so it would assign all necessary propositions a probability smaller than 1.

Yet inconsistencies feel much more troublesome when they compromise practical rationality. It is easy to ignore inconsistencies when they are immaterial to the agent's sphere of decisions. A belief is immaterial to a set of decisions when, ceteris paribus, the agent would have made the same decisions even if she did not hold the belief. When none of an inconsistent set of

beliefs is material to the agent's sphere of decisions one can always isolate them by leaving them out the psychological understanding of the agent. For example, suppose that my boss has inconsistent beliefs about a surgery that her doctor has recommended her to have, namely, she believes that it is very dangerous when told about its 20%-failure rate and also believes it is very safe when told about its 80%-success rate (Tversky & Kahneman 1974). If she and I keep a strictly professional relationship, by itself this inconsistency will not normally affect my psychological understanding of her, for neither belief is material to the sphere of her decisions in which I am interested, namely, her professional decisions. I can therefore isolate my understanding of her from the inconsistency. However, I claim that no such isolation is possible when any of a set of inconsistent beliefs is material to an agent's relevant sphere of decisions. The reason is that when this happens the inconsistent beliefs moot any maximizing interpretation of the agent's decision making.

Beliefs come into play at two moments in the decision-making process. First, they feature in the calculation of the consequences of actions given relevant conditions. When I figure that the consequence of doing A if condition C occurs is O, this both is normally inferred from other beliefs and itself constitutes a belief of mine. Secondly, beliefs feature in the assignment of probability to the conditions relevant to a decision. Once I have determined all the consequences of an action given each condition, I must weigh each of the separate utilities by the likelihood that each condition will obtain. In both cases, beliefs shape an agent's preferences among actions.

The belief system is, therefore, a determinant of the preference system. Moreover, an agent's preferences are not actual acts of choice but abstractions over her beliefs and her basic values, as implicit in revealed preference theory (e.g. Houthakker 1950, Sen 1971). A preference

system comprises preferences between actions whose comparison the agent has never entertained. It is, in short, an idealization reflecting the preferences implied by the agent's beliefs and basic values.

When an inconsistent belief that *p* is material to the agent's choice of A among a set of options, then a belief that ∼*p* would imply a preference for something else, say, B. For, assuming that the agent does not believe contradictions, had she believed that ∼*p,* she would not have believed that *p,* hence she would have chosen differently given that the latter belief was material. Since the set of beliefs held by the agent inconsistently with her belief that *p* strictly implies ∼*p,* it determines a preference of B over A. Therefore, when an agent has inconsistent beliefs that are material to her sphere of decisions, not only her beliefs are inconsistent, but so are her preferences: they imply preferences of A over B and also of B over A. And when an agent's preferences are inconsistent maximizing explanations are moot. For one thing, the preference system cannot be integrated into a utility function (Samuelson 1938, Houthakker 1950). Worse, there is no guarantee that the set of maximal options will be nonempty for all decision problems (Sen 1970, Ch. 1; 1997) and so, on the maximizing conception of psychological explanation, there is no guarantee that the agent's decisions will have an explanation.

Akrasia

Akrasia is the condition of acting against one's best judgment (Davidson 1970b). Imagine I am home by myself and the time comes for dinner. After a quick deliberation, I realize that, all things considered, my best option is to make myself a healthy salad and spare myself some extra pounds. Yet I pick up the phone and order Chinese food instead. There is a strong intuition that I have not acted rationally for I have acted against my own best judgment. Where does this intuition come from?

The answer to this question will have to involve some qualification because there is a seeming paradox associated with akrasia. Sure I acted irrationally in ignoring my best judgment, but I also acted rationally given the reluctant adoption of my intention to get Chinese food (Setiya 2007). I called the Chinese take-out, placed my order, waited for the delivery guy, paid him, etc. In short, I took the best steps necessary to fulfill my intention, which is what one would rationally do. The seeming paradox is the result of the previously mentioned vagueness associated with "rationally." As we will see the akratic action is rational according to one standard and irrational according to a stricter one.

The irrationality involved in a case of akrasia is transparently due to maximization failure. As Davidson 1970b introduced the expression, "to act against one's best judgment" does not refer to the judgment which is, in some sense (e.g. thoroughness), best compared to the others. Rather, it refers to the judgment *about* what is best to do. On the other hand, since the concept of utility is broadly defined to include all sorts of values and since expected utility is defined in terms of one's own system of values and credences, acting against one's best judgment is, in this sense, simply failing to comply with the decision-theoretic optimality condition. My proposal thus immediately explains the sense in which akratic agents act irrationally.[16]

What makes akrasia a particularly flagrant failure to maximize utility is the fact that the agent is fully aware of it. Other cases of irrationality involve, for example, calculation mistakes (of outcomes, probabilities, etc.). One cannot then rationalize the agent's action because some decisions leading to it do not satisfy the decision-theoretic optimality condition because of the mistake. In such cases, however, the agent does not normally own the action, for the mistake

[16]Arpaly 2000 has argued that akrasia need not involve irrationality but her arguments seem to obviate Williams' distinction between internal and external reasons (1979). The akratic agent may be doing what is best for her, but the problem remains: she, of all people, believes she is not!

accounts for her failure, defeating thereby the expectation that it can be rationalized. The action is a blunder.[17] But in the case of akrasia, there is no mistake. The agent willfully chooses an option with less expected utility than others she considers. She can't therefore disown her action and we are at an impasse: we cannot make sense of it nor can the performer disown it. There is nothing either party can do to correct the interpretative situation.

Akrasia is no pathology. Rather it is a commonplace, even natural, phenomenon which surely has a neural basis. Brains are complex organs which must coordinate very disparate responses to scores of conditions. These conditions oftentimes overlap and so a crucial function of the brain is to inhibit responses when they conflict, so as to produce a coordinated and coherent response to the world. In contrast, psychological explanation relies on the assumption of unity. We assume not only that whatever goes on inside the brain produces a unitary, time-coherent behavior, but crucially that this behavior (including speech in the case of humans) is the manifestation of a relatively small number of homogenous dispositions. Though this turns out to be a miraculously reliable approximation, it is an approximation nonetheless.

Even in incorporating these shortcomings, the decision-theoretic account I am proposing lives up to expectations. For nothing can be clearer than that the value and credence function are idealizations (see below). And even without appealing to such idealizations as functions, a person's system of preference is obviously an abstraction from her credences and basic sets of values while these, in turn, are approximations to her dispositions, as made explicit by revealed preference theory (Sen 1971, Ch. 1).

A monument to our every-day reliance on the approximation of unity is the concept of self-control. For one thing, the introduction of "self-control" into our language is a testimony of the imperfection of folk psychology, as it captures all sorts of situations where belief-desire

[17] It may even lose its status of an action, properly speaking.

explanations of actions yield at best confused results. But for another thing, recent research in neuroscience has shown that self-control itself does not have a unitary neural basis, but rather it is the result of the interplay of many areas of the brain having different known functions. People usually exhibit different amounts of self-control in different types of decisions and under different conditions. For example, my wife may be great at focusing on her reading in distracting environments but cannot control her emotions when she sees a spider. It turns out that the metaphor of the brain as a parliament rather than tyrant falls short of conveying the multiplicity of fronts in the struggle for dominance—unless we are talking about the current US congress and the armies of lobbyists hovering around.

If we ignore the gripping effects of emotional urges such as fear or anger, the metaphor of the parliament is not too far off the mark in demarcating the approximation of unity. As Daniel Kahneman (2011) has shown, our most ordinary problem-solving skills are best described as the interplay of two distinct systems (see above). These two systems are characterized by a number of functional properties. But, most important for the present purpose, they implement radically different problem-solving strategies. Moreover, the intervention of one (but not the other) is energy consuming, which manifests in what has recently been called *ego depletion* (Baumeister et all 1998). When tired we are more likely to rely on the subconscious, automatic system, fast-tracking decisions that would otherwise have received a different treatment. Since ego depletion cannot easily be captured in terms of beliefs and desires (or credences and values), it exemplifies the limitations of the belief-desire (or credence-value) explanation.

In fact, the idea of mental fatigue resonates with the idea of weakness of will, a phenomenon closely related to akrasia (Holton 1999, McIntyre 2006). For the metaphor of depleting our mental-energy supply suggests weakness. Experiments confirm this assessment

showing that ego-depleted individuals are much more likely to exhibit weakness of will by, for example, failing to stick to their resolutions. No wonder in extreme conditions of stress and fatigue human behavior becomes much more volatile and aggressive. But even aside from the gripping effect of emotional urges, the more reasonable of the two decision-making systems posited by Kahneman (in his terminology, System 2) is pushed aside by the more intuitive one (System 1) when individuals are ego-depleted. It is safe, therefore, to conclude that weakness of will is a deficiency of self-control; hence, often the manifestation of the struggle of various systems in the brain and, particularly, a defeat of the executive control system (see above).

In sum, phenomena like akrasia and weakness of will reflect an imperfect fit between the decision-theoretic explanation of behavior and the actual workings of the brain. On the one hand, this imperfection accounts for their categorization as instances of irrationality. This usually results in rational criticism of akratic persons by their countrymen which, in turn, translates into pressure to minimize the manifestation of these epistemic defects or to at least pack them into their private sphere. However, this folk account of the phenomena is biased by our partisanship in the social fabric. We have incentive to blame agents for these phenomena in order to create pressure toward predictability by the decision-theoretic strategy. Yet, if what I have said here is correct, the truth is not so harsh with akratic agents. A big part of the problem does not lie in them but in the decision-theoretic explanation of behavior. In fact, it seems likely that any explanation of behavior simple enough to work as the central framework of folk psychology would have to be an approximation. Still, if we are to approximate, the decision-theoretic explanation is the best approximation, given its flexibility.

As we saw, according to the decision-theoretic account I have proposed, rationalizations depend on the explanatory interest of the interpreter. At a minimum, I can rationalize an action

by recognizing the decision to perform it as optimal among all the options leading to the immediate goal or intention of the performer. But this recognition may not satisfy me if what I am looking for in the action is a rationalization of its immediate intention. In such a case, my standards of rationality are higher for I seek recognition that at least two stages of the plan comply with the decision-theoretic optimality condition. If my standards are higher, I will likewise expect compliance with the optimality condition at higher levels in the plan structure.

Oftentimes our explanatory needs are relatively superficial. When walking down a busy street I do not need to know where everybody is going in order to recognize their intention to avoid collision and act accordingly. Or when I buy a product, the seller does not need to know what I intend to do with it to understand my behavior and react in a way appropriate to her own interest. For this reason, decision-theoretic explanations can fail to subsume some decisions within a plan without losing a grip on all the actions involved. In other words, decision-theoretic explanations allow us to isolate the irrational part preventing infection of the related decisions and actions. To be sure, you cannot understand why I decided to get Chinese food while my best judgment so decidedly told me to make myself a salad. But given my intention to get Chinese food you can rationalize all my subsequent actions. From the point of view of the delivery boy, this is all that matters. His standards of rationality are understandably low and for him I acted perfectly rationally.

Self-Deception

Another folk form of irrationality is self-deception. The philosophical literature on self-deception has been divided over whether the phenomenon is best modeled after interpersonal deception. When an agent engages in self-deception, does she literally deceive herself? If so, she must have the intention to deceive herself and must believe that about which she intends to deceive herself.

This raises a host of puzzles that some philosophers have tried to address (e.g. Johnson 1988, Bermudez 2000). The alternative has been championed by Mele (2001) and consists in viewing self-deception as a type of wishful mishandling of evidence. Because it is still not obvious that the first account is at all a psychologically plausible description of garden-variety cases of self-deception, I will be partial to the second and simpler account. The proposed decision-theoretic account of rationality turns out to be a perfect match for this account of self-deception.

Mele's account is based on a psychological account of hypothesis testing inspired in signal detection theory (Trope & Liberman1996). According to this account, the processing of evidence by people is not a basic, unitary cognitive process but rather it can be broken down into tasks and decisions varying with pragmatic conditions. In particular, the costs that an individual assigns to the two possible types of errors associated with the testing of hypotheses, that is, false acceptance and true rejection, crucially determine the way she will handle the evidence. If the cost of false acceptance (accepting the hypothesis when it is false) is significantly higher than the cost of true rejection (rejecting the hypothesis when it is true), normal individuals will adopt what I will call a bias toward falsity, that is, they will approach the evidence to minimize the likelihood of false acceptance. If, conversely, the cost of true rejection is significantly higher than the cost of false acceptance, normal individuals will naturally adopt a bias toward truth, that is, they will approach the evidence to minimize the likelihood of false acceptance.

To use one of Mele's examples, imagine that a high CIA official is suspected of espionage and, accordingly, has been accused of treason. Naturally, her parents take her side and this is explained by the fact that they perceive the cost of false acceptance of the charge as very high compared to the cost of true rejection. Consequently, they will handle the evidence available against their child with a bias toward falsity. For example, they will focus on evidence

that tends to disconfirm the accusation or will interpret the evidence in ways that confirm her innocence. But for her staff of CIA agents, though perhaps they feel collegial loyalty toward their leader and so the cost of false acceptance of the charges is not negligible, the cost of true rejection might be fatal, as that would result in their continuing to operate in imminent danger. As a result, we may expect them to be much more impartial in assessing the evidence available in the case. For example, they will treat all the evidence, whether it confirms or disconfirm the accusation, with the same attention and will not gratuitously seek convoluted interpretations of the evidence when simple interpretations are available.

This theory, whose basic framework was introduced in Friedrich 1993, has changed the conventional wisdom about evidential biases. Whereas before this theory biases such as the confirmation bias or the positive re-interpretation bias were considered breaches of rationality, they are no longer viewed as inherently irrational. The change of perspective has been driven by a change in the way psychologists think of the processing of evidence. Before they thought that people sought truth and only truth with their epistemic endeavors and that evidence processing was a basic, unitary cognitive process. Now they think that dealing with evidence is just another kind of action informed by pragmatic considerations that can give priorities to one of the many aspects of evidence processing.

As I understand him, Mele views self-deception as a spurious case of evidential bias. According to Mele's account, for a subject, S, to be self-deceived about *p* roughly the following conditions must obtain: 1) *p* is false; 2) S believes that *p* based on a biased processing of the evidence; 3) this evidential bias is motivational, that is, caused[18] by S's desire that *p*; and 4) the body of data available to S at the time provides greater warrant for ~*p* than for *p* (2001, pp. 50-51). While this analysis seems to me to be on the right track, I disagree with Mele's claim that

[18] In a non-deviant way. See Mele 1992.

(1)-(4) are jointly sufficient for self-deception. Imagine that in the current tough market conditions, I get an email saying that a philosopher is needed immediately who can teach an undergraduate Bioethics course at an excellent school, with more than competitive salary and excellent research conditions. Bioethics is not my area of specialization, nor have I ever taught it. But I have read a few articles about it here and there, and have had some interest in the subject. In all, imagine that the evidence available to me supports my inability to teach a course on Bioethics but that it is nonetheless a close call. I am faced with the decision to respond to the email which in turn depends on accepting or rejecting the hypothesis that I am able to teach an undergraduate course on Bioethics. Suppose that upon considering the evidence with a bias I conclude that, with some effort, I would be able to teach the course and so decide to respond the email. Did I engage in self-deception? It seems to me that I did not. In fact, it seems to me that I may have done the rational thing because missing such an opportunity based on an error of judgment would be catastrophic for my career. Yet I satisfy all of Mele's conditions: I believe falsely that I am able to teach Bioethics, I have such a belief based on a biased processing of the evidence, this evidential bias was caused by my desire that I can teach Bioethics (so that I can take the position), and the evidence objectively provides greater warrant for the opposite conclusion.

If the thesis I am defending here is correct, decision theory should shed light on what makes self-deception a form of criticizable irrationality. In fact, once we trace the linage of Mele's theory to signal detection theory, this should be unsurprising. Understanding evidential biases in terms of decision theory sharpens Mele's account and reveals the true nature of self-deception.

The key lies in recognizing that hypothesis testing is, usually, an action. I will exclude

from this claim perceptual hypothesis testing since the fact that we are prey to visual illusions even when we know them to be illusions complicates their inclusion. But there are plenty of cases of hypothesis testing that are under voluntary control and actively pursued, such as the testing of scientific hypotheses. In fact, hypothesis testing occurring in the early stages of perception may be the only one that is involuntary, all other hypothesis requiring the willful investment of intellectual resources. If so, we can understand the action of testing a hypothesis and, particularly, the adoption of a bias in so doing, in light of the maximizing conception of rationality. The exercise requires the resolution of some technical complications, however.

First, acting on true beliefs usually leads to success in realizing a world that one values. It is, therefore, usually rational to seek the truth in the purely instrumental sense that believing the truth tends to increase one's utility. Agents do not seek to maximize their *expected* utility, they seek to maximize their utility. It is due to the fact that they cannot escape their epistemic limitations that they have to settle for expected utility as the best available approximation to their utility given their uncertainty (see chapter 5). But when the fixation of beliefs themselves is the object of their deliberation it is easy to realize that the closer the beliefs are to the truth the more likely agents are to succeed in their plans and, consequently, the more utility they are expected to attain from their actions. Hence, according to the decision-theoretic framework, it is rational to have a basic drive to incorporate into one's belief system (in Bayesian terms, conditionalize on) new evidence, for new evidence cannot but decrease the statistical distance to the truth—provided the evidence is itself not tarnished by falsity. In fact, such a drive is generally observable.

Secondly, like any other decisions, we can model decision-making about how to seek out and handle evidence with decision theory. This sheds new light on Friedrich's analysis of biases

in terms of the costs of the two types of error. Given an agent S epistemically interested in a proposition *p,* S will be faced with the options of adopting any bias in the cognitive process leading to settling on P's truth-value. The biases may affect not only the selection and processing of the evidence but also the selection of the hypotheses to be tested. For example, S may choose to probe only the hypothesis that *p* or she may also be attentive to evidence that tends to confirm $\sim p$. In fact, S might contemplate alternative hypotheses compatible with the evidence in trying to determine whether *p* is the best. However, in making the decision to adopt any biases the subjective probabilities associated with the various relevant states are a priori, in that they must be independent of the evidence being processed.

The scenarios that can result from inquiring into *p* are: accepting *p* and *p* being true ($Ap \cdot p$), accepting *p* and *p* being false ($Ap \cdot \sim p$), rejecting *p* and *p* being true ($Rp \cdot p$), and rejecting *p* and *p* being false ($Rp \cdot \sim p$). Hence, the utility associated with a cognitive endeavor, E, to increase one's degree of confidence in *p* is:

$$U(E) = U(Ap \cdot p)P(Ap \cdot p|E) + U(Ap \cdot \sim p)P(Ap \cdot \sim p|E) + U(Rp \cdot p)P(Rp \cdot p|E) + U(Rp \cdot \sim p)P(Rp \cdot \sim p|E).$$

To simplify notation, let $U(Ap \cdot p) = U_1$, $U(Ap \cdot \sim p) = U_2$, $U(Rp \cdot p) = U_3$, $U(Rp \cdot \sim p) = U_4$, $P(Ap \cdot p|E) = P_1$, $P(Ap \cdot \sim p|E) = P_2$, $P(Rp \cdot p|E) = P_3$, $P(Rp \cdot \sim p|E) = P_4$. We thus have:

$$U(E) = U_1P_1 + U_2P_2 + U_3P_3 + U_4P_4.$$

Basic Bayesian algebra shows that $P_1 + P_3 = P(p)$ and that $P_2 + P_4 = P(\sim p)$, and so, as expected, that $P_1 + P_2 + P_3 + P_4 = 1$. $P(p)$ and $P(\sim p)$ represent the a priori subjective probabilities the agent associates with p and $\sim p$, respectively, which are independent of the adoption of any cognitive endeavor, *E*, and so are constants for our purposes. In contrast, P_1, P_2, P_3 and P_4 depend on *E* and so the agent has control over their relative weights by choosing one

cognitive endeavor or another. Moreover, as Trope and Liberman (1996) correctly define it, the cost of false rejection (viz. type I error) is the loss that results from rejecting a true hypothesis compared with accepting it, while the cost of false acceptance (viz. type II error) is the loss that results from accepting a false hypothesis compared with rejecting it (1996, p. 252). Calling them C_I and C_{II}, respectively: $C_I = U_1 - U_3$ and $C_{II} = U_4 - U_2$. Again with the help of some simple algebra, we arrive at the following:

$$U(E) = U_1P(p) + U_4P(\sim p) - (C_IP_3 + C_{II}P_2).$$

We can see the negative term as a Bayesian average of the costs of false rejection and false acceptance; something like the expected cost of adopting a cognitive endeavor, *E*. The positive term, by contrast, is a constant with respect to *E*. Consequently, decision theory mandates that when various kinds of cognitive endeavors are available we choose the one that simply minimizes the expected cost.

Consider now the choice between two cognitive endeavors: adopting a bias toward truth and adopting a bias toward falsity. Clearly, a bias toward truth will result in a smaller P_3 at the expense of a larger P_2 and vice versa. Hence, since these are the factors pondering the costs of false rejection and false acceptance, respectively, we can think of the choice as the decision which of the two costs one wants to control in order to minimize the overall expected cost. Obviously, if the cost of false rejection is larger than the cost of false acceptance, one will maximize her utility by minimizing the impact of C_I in the expected cost, hence by adopting a bias toward truth. If conversely the cost of false acceptance is larger one will maximize utility by minimizing the impact of C_{II} in the expected cost, hence by adopting a bias toward falsity. This is exactly what Trope and Liberman (1996) concluded using only folk psychology and common sense.

Trope and Liberman also observe that whether one should at all inquire into *p* depends on the cost of inquiring or, as they call it, the cost of information (1996, p. 253). Once again, this is predicted by the decision-theoretic model.

The cost of information is just the opportunity cost or the utility that is forgone by inquiring into *p*. According to standard economic theory, this corresponds to the maximum utility the agent could have attained by using her time and resources in other ways. Assuming that guessing is a trivial form of inquiry demanding no investment of time and other resources, it can always accompany any other endeavor and so we can count on the free enjoying of at least $U_1P(p) + U_4P(\sim p)$, as this term in the above formula is independent on the choice of cognitive endeavor. Hence, the utility gained by non-trivially inquiring into *p* (viz. not guessing) instead of pursuing other endeavors is simply the decrease in the expected cost resulting from it. When the reduction in expected cost resulting from non-trivially inquiring into *p* outweighs the utility resulting from the best alternative endeavor, we should non-trivially inquire into *p*; otherwise, we should not. Obviously, the more time and effort it takes to reduce P_3 and P_2 so as to reduce the expected cost a unit, the more utility one could attain by alternative endeavors, and so the less profitable it is to inquire into *p*. In other words, the higher the cost of information the less beneficial it is inquiring into *p*. There must therefore be a threshold beyond which we should not, as Trope and Liberman claim.

Let us now return to self-deception. As we can see, in general there is nothing irrational about inquiring into *p* with a bias toward acceptance or rejection. This is so even when it happens that the agent desires that *p* or that $\sim p$. So, pace Mele, this cannot be what constitutes self-deception for self-deception is typically irrational. What makes biases toward acceptance or rejection rational is simply the difference between the cost of false acceptance and the cost of

false rejection. The Bioethics-course case mentioned above shows that the cost of false rejection can be much larger than the cost of false acceptance even when the agent desires that *p*. The reason why biases can be rational is that when the cost differential is big enough the expected utility resulting from a biased inquiry can exceed the expected utility of conducting an unbiased inquiry. I will now suggest that the problem with self-deception lies in what counts as an increase in utility.

I can increase my expected utility in two ways. I can act upon the world so that to change it in a way that realizes more of my values. This is what we normally try to do. But also I can alter my credences (or beliefs) so that the degrees of confidence I have in the events that I value positively increase relative to those I value negatively, thereby boosting up my expected utility quite generally. However, it is hard to avoid the feeling that the last way is cheating. Worse still, it is cheating oneself. Of course, altering my credences in this way voluntarily and without even the pretense of appealing to evidence would require a form of self-manipulation that raises the same kinds of puzzles that have saddled the interpersonal models of self-deception. For such an action would require that I have the intention to change the beliefs I know myself to have. However, Trope and Liberman's account of evidential biases provides an alternative, psychologically healthier avenue to achieve the same result. I may systematically adopt acceptance biases toward propositions that I value and rejection biases toward propositions that I do not, changing thereby my degrees of confidence in them in the way required. I would not be forming beliefs contrary to those I already have but instead I would simply be forming new beliefs or altering those that I have. Hence, no puzzle about beliefs comes up. Also, since the strategy is piecemeal and opportunistic no intention to deceive myself is necessary. Instead, it only requires epistemic gullibility.

Yet, all the same, I would be the only victim of such an epistemic policy because, by altering my credences in the way described, I would put myself in a worse position to bring about a world I value. True, I will have increased my expected utility, yet not by increasing my utility but by impoverishing my belief system. Because utility, not expected utility, is what agents ultimately seek to maximize, altering my beliefs in a way that responds neither to evidence nor to utility but only to expectation would be a monumental failure of rationality and could quickly lead to catastrophe. Fortunately, people rarely are so spectacularly irrational as to altogether miss the fact that the expected utility is only a guide to the ultimate goal, namely, utility.

But people often achieve the same effect in a subtler way. A widespread example is altering one's credences not simply to match one's values but to avoid an emotional burden. In fact, it is widely accepted among social psychologists that most people are optimistic in that they assign more credence to states they value (e.g. Kahneman 2011). The explanation that they do so in order to increase their expected utility by impoverishing their belief system seems to me too morbid to be applicable to most people. Rather, it is natural to think that they value a world where they are not under stress and this translates into a cost of false rejection of the propositions they value biasing thereby their cognitive processes. There would seem to be nothing irrational about factoring in emotional costs into one's utility, for these definitely make a world of difference to our happiness and satisfaction. Yet oftentimes it is.

With the development of neuroscience we have acquired a much better understanding of emotions. Although there still is much disagreement on a number of issues, there is some consensus that emotions play an important role in decision making (de Sousa 1987, Damasio 1994, Prinz 2004). In deliberating, emotions serve to appraise options based on similar past

experiences. This allows organisms to integrate their past experiences into their present deliberation in an efficient way (e.g. Bechara, Damasio & Damasio 2000). Individuals with brain lesions that affect emotional processing usually exhibit deficiency in decision making (Bechara, Tranel & Damasio 2000). It seems quite plausible indeed that the advantage that the proper working of emotions provides for decision making in animals is one important force that drove its evolution.

Consider the case of fear. Fear is an emotion associated with the release of hormones that result in an increase in alertness, responsiveness and, generally, trigger a kind of self-preservation mode. These kinds of responses affect decision making in ways that can be crucial for the survival of the organism. It is believed that fear, therefore, evolved because of the survival advantage that such an alteration in decision making can have in dangerous situations.

But if the function of emotions is to positively influence decision making, viewing one's experiences of emotion as objects of intrinsic value amounts to focusing one's decision making on decision making rather on the world. Emotions are tools of decision making, not its target. In deliberating, the fact that we experience an emotion in a hypothetical state of affairs should not by itself add or subtract any value or utility to the state of affairs' evaluation, for emotions are supposed to co-instantiate with states of affairs that are good or bad precisely so that we can better react to them, not to make them better or worse. Of course, we generally evaluate negatively scenarios where we can anticipate we will experience negative emotions. But this is not due to the presence of the negative emotion but rather the other way around, namely, the presence of the negative emotion should hint us to badness of the scenario. Emotions are mere autonomic signals that help to efficiently appraise worlds as good or bad; they do not constitute the worlds' goodness or badness.

True, people ordinarily say that they long for emotional balance or that they want to be in control of their emotions. But they mean either that their emotions are not working properly and need a little tuning, or that they have been going through a lot difficulties which, in turn, has led them to experience a lot of negative emotions. Normally what they mean is that they need to fix their lives by making decisions that are going to lead to a world that they value so that they do not experience so much negative emotion. In fact, I suppose they would not be content with going through all the difficulties without experiencing any negative emotion, for that would involve a kind of insensitivity to difficulties that could only worsen their reactions to them (Bechara, Tranel & Damasio 2000).

I claim that self-deception consists in adopting an evidential bias as a result of assigning an intrinsic cost to a burdensome emotion. In other words, self-deception occurs when the (so-called) emotional cost of acquiring a belief that p alters the balance between C_I and C_{II} leading her to bias her inquiry into *p.* The characteristic irrationality of self-deception derives from the fact that the agent has failed to maximize her utility in the genuine emotion-insensitive sense of the word. But, moreover, by doing so she has blindfolded herself to the world for no benefit. She has not taken the world as the target of deliberation and so she has temporarily decided to ignore what is most crucial for her future well-being. All that self-deceived agents accomplish, on this account, is shielding themselves from bad news, but the news is no better for that. To borrow an expression from David Lewis (1981b), playing the ostrich is never rational. Although this analysis is not based on the interpersonal model of *deception*, I think this self-inflicted misinformation is what accounts for the aptness of the word to describe the phenomenon.

The idea that emotions are involved in self-deception is not new (Bach 1981, Johnston 1988, Barnes 1997). But their exact role has remained obscure. For example, in a similar vein as

Mele, Johnston 1988 and Barnes 1997 talk of self-deception as connected with a mishandling of evidence motivated by an "anxious desire" that the relevant proposition is false. My proposal sheds light on this point. It is not just that the desire is accompanied with anxiety, as I may anxiously desire that I am qualified to teach the Bioethics course, accordingly adopt an evidential bias yet not be self-deceived if I am justified in doing so. It is the fact that the anxiety factors in my evaluation of my options which constitutes the failure of rationality. The problem is that because I do not want to feel anxious or discouraged or ashamed (which I anticipate in my deliberation I would feel if I believed the relevant proposition) I adopt a bias toward falsity. Yet in doing so, I have made myself less likely to find out the truth for no gain in expected utility.

For example, cases of self-image are frequently mentioned in connection with self-deception. On this account, when the reason for, say, positively misinterpreting the evidence against oneself in a way that allows one to retain her self-image is fear of the disappointment, discouragement, etc. that such an acceptance would bring with it, one has retained her self-image through self-deception. Similarly, if the reason why the parents of the CIA official took a bias in inquiring into the accusations of treason against her specifically because they wanted to avoid the pain and shame of accepting that their child is a traitor, they are self-deceived about their child's innocence.[19]

However, I should make it clear that it is not always irrational to incorporate our emotional states into our evaluations of various options in decision making. What is irrational is assigning them intrinsic (non-instrumental) value. Consider the case of a pitcher who is making his debut in the major leagues. He is brought into the game in the 7th inning and asked to get through that inning. After three batters faced, he is yet to get an out: he has allowed a hit, hit

[19] We may or may not want to restrict self-deception to cases in which the relevant belief is false, as Mele does. I am not sure if this comports with our folk use of the term. Mele is under more pressure to include this condition because his general account of self-deception is exceedingly permissive. My account does not have that problem.

another batter and walked the third one. At this point, there is plenty of evidence that he is not a good pitcher, evidence that the team's manager must be particularly aware of. However, believing that he is a terrible pitcher at that moment when he is really not would surely kill his morale and lead him to perform even worse. If only to avoid giving up runs, it would seem rational of him to adopt a bias toward rejection of the hypothesis that he is a terrible pitcher. If he manages to believe that he is the best, chances are he can get out of the jam. Here, however, he would not do it to avoid the emotion for its own sake but to avoid it in order to bring about a world he values. Probably he has what it takes to be in the majors, for he has proved to be in charge of his emotions.

Whether this case counts as a rational form of self-deception or whether we should tie the expression to irrationality, I do not know. But this explains why normal people have a tendency to be optimistic without attributing irrationality to everybody. Sure they want to avoid the emotional burden of bad news, but many times the reason for wanting to avoid it is that they fear such a burden would jeopardize their performance in their social setting or professional environment. Whether or not they are justified in having such a belief is not the point. The point is that they are all the same exonerated from the charge of irrationality because they do not want to avoid the emotional burden for its own sake.

Chapter 4

INTENSIONALITY AND FOLK PSYCHOLOGY

Intensionality is often considered one of the marks of the mental (e.g. Chisholm 1957, Ch. 11). This is usually presented as the fact that substitution of co-referentials somehow fails in sentences reporting beliefs and other intentional mental states. Yet, for intensionality to be a mark of *the mental*, the failure of substitution must not be a purely linguistic phenomenon but rather must reveal something about the mental states. What exactly? Here there is no consensus and instead much obscurity. In this chapter I argue that the claim that mental states, as opposed to sentences reporting mental states, are intensional leads to serious difficulties that we'd better avoid.

I. Intensionality and Folk Psychology

Historically, the intensional/extensional distinction has been understood in relation to language (Frege 1892, Carnap 1956a, etc.). This is unduly restrictive, however. In its broadest sense, languages are compositional systems of symbols. To be sure, natural languages such as Spanish or English can be intensional (or extensional). In fact, the distinction presumably applies to languages whose function is not at all communication. For example, computer languages seem intensional but their function is to serve as a medium for the realization of the computer's activity rather than communication. But I don't know of any reason why non-compositional systems of representations cannot be intensional (or extensional). In the absence of reasons to the contrary, I will take the intensional/extensional distinction as applicable to any system of representations.

The most widely accepted criterion for intensionality is the failure of substitution *salva veritate*. Other criteria have been proposed for the intensional/extensional distinction, including

the existential generalization test. However, it is not clear to me that these coincide in all cases (see chapter 7). Still, the substitution of co-referentials *salva veritate* subsumes in large measure the cases that have motivated the introduction of the distinction such as sentences of identity. But these cases were concocted to illustrate a deeper point and the failure of substitution of co-referentials *salva veritate* is too strong a condition in several respects.

First, "reference" is a word normally used for singular terms (in Kaplan 1977's sense) as opposed to predicates. Yet there is no reason to rule out the possibility that a representational system is intensional (or extensional) in virtue of the substitution of predicates. Hence, it will be better to speak more neutrally of substitution of *representations of the same thing*, instead of *co-referentials*.

Second, there are systems of representations which do not have truth-values or whose truth-values are irrelevant to their function. For example, the truth-value of a sentence in a computer language is quite immaterial to its function and so the language's intensionality would seem irrelevant to understanding it. This rings counterintuitive. In fact, in paradigmatic cases of intensionality, such as sentences of identity or belief reports, it is not the failure to preserve truth-value that makes them intensional, but rather the failure to preserve what they mean or communicate (chapter 7 below; Kripke 1979, Crimmins & Perry 1989, Bach 1997). Since a difference in truth-value guarantees a difference in meaning, the appeal to truth-value has been used to make the cases forceful; but it is not an indispensable mark of such cases.

A general definition of intensionality/extensionality must therefore make reference to the function of the representational system rather than truth. More precisely, it must depend on the representational function of the system, that is, the function that its representations have in virtue of representing what they do. Communication of information about the particular aspects of

world surely is the primary representational function of natural languages such as English and Spanish—hence, the definition of intensionality as failure of substitution *salva significatione* (Kripke 1979). But in the case of a computer language, its representational function is to modulate the computer's activity and its intensionality should accordingly be traced to the fact that two symbols of the same thing may have different computational roles and potentially (though not necessarily) substitution of one for the other may result in a different outcome of a computational process in which it participates.

In sum, I will say that a system of representations is extensional iff representations of the same thing always have the same representational function; otherwise, it is intensional.[1] In a nutshell, a system of representations is extensional iff substitution of two representations of the same thing never makes a difference.

Because whether two representations represent the same depends on how we individuate what they represent, so does whether a representational systems is intensional or extensional. A notorious example is explanations. Since ancient times philosophers have been debating whether a predicate represents a property (or universal; Armstrong 1978) or a set of (actual) objects (e.g. Lewis 1983b). These objects are those to which, intuitively, the predicate applies and form a set usually called the predicate's *extension*. The name is unfortunate for my purposes since, even if predicates represent their extensions, sentences in which they occur may still be intensional in the sense that interests me here. In fact, this might be the case with explanations. If explanations are answers to why-questions (van Fraassen 1980) and if, presumably, answers to why-questions are types of utterances (including, in fact, types of sentences) individuated by their meaning, then they are, broadly speaking, representations. Hence, explanations are the kinds of things that can

[1] Neale 1990 argues for a three-fold distinction where a non-extensional context can still be non-intensional by being what he calls (following Cresswell 1975) *hyperintensional*. The only hyperintensional contexts I know are sentences reporting intentional states such a belief, desires, etc. I will address these in chapter 7, but my definition here does not deny hyperintensionality but simply clusters it with intensionality.

be extensional or intensional. But if we take predicates to stand for their extensions then explanations are intensional since clearly that X has a heart is an explanation for why there is blood distributed all over her body, yet that X has a liver is not, even though "has a heart" and "has a liver" are predicates with the same extension (cf. Levin 1979). If, on the other hand, predicates represent properties, then, though co-instantiated, surely the property of having a heart and that of having a liver are distinct; hence, the predicates "has a heart" and "has a liver" are representations of different things; hence, substitution of representations of the same thing does not fail in the above example. Explanations may still be extensional—in our sense.

Historically, the problem with treating predicates as representing properties has been the latter's metaphysical respectability. Yet with the development of modal logic, this obstacle has been, in good measure, overcome by treating properties as functions from possible worlds to sets of objects, identified with the corresponding predicates' extensions at a given world (e.g. Carnap 1956a). This sits well with recent accounts of explanations since they seem to be intimately connected with counterfactuals (e.g. van Fraassen 1980, Lewis 1986a). Provided that substitution of co-referential expressions preserves explanatory role, we can thus view explanations as extensional in our sense if we view predicates as representing functions of possible worlds to sets of objects.

One might think that folk-psychological explanation must be an exception since the way we express the content of a mental state makes a difference to psychological explanations. For example, that Lois Lane believes that Superman is handsome is an explanation of her flirting behavior toward him, even though that Lois Lane believes that Clark Kent is handsome is not, because she believes no such thing. Yet "Clark Kent" and "Superman" are representations of the same person and so substitution of one for the other seems not to preserve the explanation. Thus,

one might conclude, psychological explanation is intensional, after all. But this would be a rash conclusion. Natural languages are sophisticated systems of communication and the assumption of compositionality does not always hold. Sometimes what a sentence communicates is not fully determined by the words that compose it. In particular, in sentences reporting beliefs, such as "Lois Lane believes that Superman is handsome," there is a wide consensus that what is communicated exceeds what results from the combination of its composite words taken in their literal meaning (see chapter 7). If so, substitution of "Clark Kent" for "Superman" in the above explanation alters the belief that is represented by the explanans and, accordingly, the case does not show that psychological explanation is intensional.

Still, if not psychological explanation, the conventional wisdom is that its explanans are intensional objects. The standard view is that beliefs, desires, concepts, ideas, etc. are intensional objects because not only do they represent things but, moreover, they can do so in different ways. However, this presupposes that they are a particular type of representation, namely, *symbols*.

I will call a symbol any representation whose individuation is not exhausted by its representational properties. The folk notion of a symbol overlaps with this definition since, for example, a dove may be a symbol of peace even though it is not individuated as a dove by its representing peace. But the folk notion is stronger since we usually think of symbols as arbitrary representations, their representational properties being entirely independent of their relevant individuation conditions. Languages, for example, are systems of symbols because words are individuated morphologically and fully independently of their semantic properties: synonymous are different words and an ambiguous term is not many words on account of its several meanings. In contrast, as I have defined it, it suffices for something to be a symbol that its

representational properties are not all that matters to its relevant individuation conditions. This is supposed to exclude things like a statue of Napoleon since whether or not a statue is a statue of Napoleon is fully determined by whether or not it represents Napoleon. I will call these representations *enactors* to distinguish them from symbols.

Obviously, enactors cannot form an intensional system of representations since two enactors of the same thing are *ipso facto* identical (that is, belong to the same representational type) and so cannot have different representational functions. Consequently, the view that beliefs, desires, concepts, ideas, etc. are intensional implies that they are not enactors, but symbols. I claim that folk psychology contradicts this implication.

II. Fodor's Trilemma

Fodor's views about the mind give rise to a trilemma that he discusses in *The Elm and the Expert* (1994). Perhaps the most distinctive of Fodor's views is *computationalism* (1975). Computationalism straightforwardly implies that the laws of psychology must be implemented in a system of formal representations in the brain much in the way programs are implemented in computers. The metaphysical substratum of the laws of psychology, namely, the lower-level facts that render them effective, consists of transformations of physical representations in the brain dependent only on their syntactic (viz. non-semantic) properties. This is the first horn of Fodor's trilemma. On the other hand, Fodor has also endorsed the so-called *informational semantics*, that is, the view that the intentional content of mental states is constituted by the information they carry about the world and, hence, depends on the relations of the agent to her environment (Fodor 1987, Dretske 1981). Informational semantics is, therefore, a form of externalism about mental content (Putnam 1975, Burge 1979, 1986). This is the second horn of Fodor's trilemma. Finally, perhaps less characteristically, Fodor also holds that the laws of

psychology are eminently intentional (with a "t"), that is, they subsume mental states in virtue of their intentional content. This is the third horn of Fodor's trilemma.

Put the last two horns together and the trilemma rises to view: if the laws of psychology subsume mental states in virtue of their intentional content and intentional content is the information the states carry about the world, then the laws of psychology tell about the agent's relations to the world. However, according to the first horn, the laws of psychology are implemented in the brain by computational (viz. purely syntactic) processes, hence independently of the agent's relations to the world. Can the laws of psychology be effective in virtue of processes that they are not about? Only if the facts that make them true (viz. agent-world relations) are always concomitant with the facts that make them work (viz. computational processes in the brain). It turns out they aren't.

Fodor's counterexample appeals to his favorite candidate for psychological law:

L) If S wants that A and believes that not-A unless B, then ceteris paribus S will perform an act intended to bring about that B.[2]

Apply L to the case of Lois Lane. According to computationalism, Lois Lane has two distinct representations in her brain, SUPERMAN and CLARK KENT,[3] which play distinct roles in her computational processes. SUPERMAN, and not CLARK KENT, crucially appears in her decision-making leading to idolizing Superman (or Clark, as one may prefer). On the other hand, if L is in fact intentional, "A" and "B" must represent the intentional content of S's mental states; hence, according to informational semantics, they must represent states of affairs, as it were, in the world as opposed to in S's brain. If so, L does not apply to Lois Lane: she satisfies the antecedent, namely, she wants Superman to love her (A) and believes that Superman will not

[2] As discussed in chapter 2, wanting that A is compatible with having no intention to bring about that A. But this is beside the point here.

[3] Here I will follow standard terminology in using upper case for computational representations in the brain.

love her (not-A) unless she spends time with him (B); yet she violates the consequent as she performs acts intended to avoid him (not-B).

Computationalism exacerbates the problem because it construes mental representations as arbitrary symbols, namely, entities whose relevant individuation conditions are entirely independent of what they represent. A computer hooked up to different sensors and actuators would produce different behavioral outputs and so the way the integrated machine embeds in the world is independent of the computing function of its inner symbols. But if mental representations are individuated independently of how the agent embeds in the world, their representational properties must be irrelevant to their individuation. Computationalism, therefore, severs the connection between mental representation and its representational target, leading to the above dilemma. Worse, no causal theory of reference or other sophisticated account of the representational properties of mental symbols (Fodor 1987) can alter the fact that the function of a symbol in a computer language is exhausted by its syntactic (hence, non-intentional) properties. The semantics of a computer language will always be a mere appendix.

But less radical accounts of mental representation still face the problem. Fodor's trilemma is an idiosyncratic version of an underappreciated dilemma stretching widely throughout the philosophies of mind and language. Cleared up of Fodor's commitments to informational semantics and the existence of laws of psychology, the dilemma is as follows. On the one hand, we seem to have no choice but to ascertain the intensionality of the mental lest we can't make sense of the fact that we obviously understand Lois Lane's psychology. But, on the other hand, we have strong reasons to believe that not only is folk psychology intentional, but that mental representations are enactors. Yet, as we saw, if mental representations are enactors, the mental cannot be intensional.

In the philosophical literature, the pressure to account for the fact that we psychologically understand Lois Lane is channeled through some classic linguistic puzzles. We unmistakably consider the first of the following sentences true and the second false:

1) Lois Lane loves Superman,
2) Lois Lane loves Clark Kent.

It seems to follow that we recognize two distinct (in fact, quite opposite) mental states associated with the same individual and these states are crucial to our understanding of Lois Lane. But if two mental states with the same intentional content can nonetheless differ, then mental states must be individuated by more than just their intentional contents, making them symbols rather than enactors. Frege was the first one to follow this reasoning leading him to postulate senses as the extra determinant of mental states. Frege's senses have fallen in disrepute but philosophers have cursorily followed him in choosing to save this horn of the dilemma at the expense of the hypothesis that mental representations are enactors. Computationalism is a more extreme manifestation of this preference by altogether denying intentional content any involvement in the individuation conditions of mental representations, thus making them arbitrary symbols.

But, despite widespread neglect by philosophers, the reasons to hold on to the second horn of the dilemma—that is, to construe mental representations as enactors rather than symbols—are no less compulsory. For, mental representations in the sense relevant here are the explanans of folk-psychology, namely, beliefs, desires, concepts, ideas, etc. And, it turns out, folk psychology individuates beliefs, desires, concepts, ideas, etc. by their intentional content only.

To begin with, the distinction between symbols and enactors is vaguely reflected in our use of "of" versus "for" in relation to representations. When symbols are viewed as

metaphysically independent of what they represent we tend to use the word for as in "'Dios' is a word *for* God in Spanish" or "does your country have a traffic sign *for* winding roads?" In contrast, when talking about representations individuated by their representational properties we tend to use "of" as in "I have the signature *of* John" or "is there a statue *of* Napoleon in Paris?" The idea is that "for" suggests a more or less free matching between representation and target because what counts as the same representation is independent of its target, while "of" suggest no such freedom. As it turns out, we would usually say that Stephen Hawking and I both have the concept *of* quark, rather than that each has a concept *for* quark, even though surely we think of quarks very differently.[4]

Consider the role of this form of individuation in our every-day lives. When we disagree about something we feel pressure to resolve our disagreement and, accordingly, when the conditions are right, engage in debate to find a consensus. Presumably this pressure stems from the logical necessity that one of the disagreeing parties has false beliefs. But if beliefs are symbols rather than enactors, there is no longer logical tension in disagreeing about something. The disagreement can always be due to the manner of representing the world rather than to there being only one world (cf. Salmon 1989).

Secondly, there is nothing else in our repertoire of folk-psychological notions to distinguish between beliefs, desires, concepts, ideas, etc. with the same intentional contents. If Stephen Hawking and I both have a belief that quarks are subatomic particles, what is there to distinguish them as of relevantly different types? Senses, modes of presentation or the like spring to mind; but these are philosophers' postulates whose unique motivation is accommodating the first horn of the dilemma and certainly not folk-psychological notions. Folk psychology

[4] It is so natural, in fact, to individuate folk-psychological representations by their intentional *contents* that in their attempt to build in further individuation conditions, some philosophers have stretched the label to include properties that are explicitly non-intentional, as witness *narrow content* (e.g. Fodor 1987, 1991). That their choice of name for their proposal is inspired in the very negation of its substance attests to the perversity of such use of "content."

developed much before men knew the first thing about the inner workings of the human body; in fact, much before they had the concept of biology. It is, therefore, highly implausible that folk psychology evolved to somehow track what happens inside the brain. Yet all there is outside the skin is the world, that is, what beliefs, desires, concepts, ideas, etc. are supposed to represent. It is, in sum, natural that these are fully individuated according to what parts of the world they represent.

This is the basis for the view, well-liked among philosophers, that propositional attitudes are dyadic relations between an agent and a Russellian proposition, in stark contrast with Fregean propositions. For, on Frege's view, propositions are composed of senses whose individuation is not exhausted by their referents and so are symbols in the sense introduced above. On a Russellian view, in contrast, propositions are complexes of objects and relations and so are not even amenable to the intensional/extensional distinction: they are states of affairs themselves rather than representations thereof. Since beliefs and desires are dyadic relations between an individual and a proposition, the state of affairs that a belief or desire represents is, in a crucial sense, constitutive of it. The relation between beliefs and desires and what they represent is thus much more intimate than other theories of the mental would have it.

A third reason why the individuation of folk-psychological representations should be enactors is simplicity. If beliefs, desires, etc. were symbolic states, psychological explanation would have an extra degree of freedom aside from intentional content. This would double up the amount of information that we must manage in psychologically interpreting others which would be surprising because the most obtuse of people are remarkably good psychological interpreters. The surprise might be dispelled if we found out that, double though it is, it is not much at all. But I have shown that a decision-theoretic framework to understand psychological interpretation is

highly complex (chapter 3) and, as I will shortly argue, decision theory is perforce non-intensional.

But perhaps the best reason to regard folk-psychological representations as enactors is explanatory adequacy. Most decisions people make would be incomprehensible if we did not allow for degrees of beliefs and desires. On the other hand, the only conceivable way to allow for degrees of beliefs and desires in psychological interpretation is by allowing subjective probabilities. Yet the incorporation of probability into psychological interpretation is incompatible with the view that beliefs and desires are symbolic states, as I shall now argue.

III. The Non-intensionality of Probability and Decision Theory

Any acceptable interpretation of probability must define it over events. Traditionally, sets of events, which we may think of as states of affairs, have been viewed as the arguments of probability functions. Since events are not representations but the very things represented by sentences, propositions, etc. they are not the kind of things that can be intensional. Nor can states of affairs be intensional, as sets are also extensional, in that membership fully determines identity. For example, in the standard Kolmogorov formulation of probability theory, if a set of events A is a subclass of a set of events B, then —never mind how you represent these sets—B must be at least as probable as A.

If we prefer to define a probability function as taking propositions as argument, we may—so long as we retain the fundamental interpretation of the function in terms of events. Many functions that are not probabilities (e.g. length, area, etc.—suitably normalized) satisfy the Kolmogorov axioms. Should we give up the association with events, little reason would remain to call any measure function a *measure of probability*. The usual view of probability as taking a proposition as argument requires, therefore, that the value of the function be the propensity,

frequency, or what you will, with which the proposition is true (e.g. Jeffrey 1983). Even if we adopt a Fregean view of propositions as representations of states of affairs, a proposition being true is intimately connected with the state of affairs it represents and so the reference to a representation is merely nominal. Given that the truth value of a proposition depends on nothing but the state of affairs the proposition represents, a probability function should depend on nothing but the state of affairs its argument represents. Hence, probability functions do not take representations as arguments and so cannot be intensional (or extensional) objects.

We can arrive at the same conclusion by viewing propositions as sets of possible worlds (Lewis 1986c). Again consider the Kolmogorov result that if a set of events A is a subclass of set of events B, then B is at least as probable as A. If we want to consider propositions as arguments of probability functions, there seems to be only one way of translating this result: if p is a logical consequence of q (viz. if the truth-condition of p include the truth-conditions of q) then p is at least as probable as q. If p and q are logically equivalent, then p is a logical consequence of q and so q is a logical consequence of p; hence p is at least as probable as q and q is at least as probable as p. Consequently, their probabilities must be identical. A possible-world treatment of logical consequence has it that p and q are logically equivalent iff they are true in exactly the same possible worlds; hence, iff they have the same truth-conditions. Therefore, if p and q are truth-conditionally identical, they must be given the same probability regardless of how they represent their state of affairs.

The expected utility principle of decision theory, in conjunction with the non-intensionality of the measure of probability, entails that the utility function cannot take representations as arguments and so it cannot be an intensional object either. Suppose we have

two truth-conditionally identical actions, h and m.[5]The principle of expected utility says,

$$1) u(h) = u(h\&x1)P(x1/h) + u(h\&x2)P(x2/h) + \cdots + u(h\&xn)P(xn/h)$$

for any partition of the space of possibilities $[x1, \ldots, xn]$ (viz. disjoint and exhaustive) and where $h\&xi$ stands for the outcome of performing h when xi obtains or, in other words, the relevant consequence of $h \cdot xi$. Consider the partition $[h \leftrightarrow m, \sim(h \leftrightarrow m)]$—where "$\leftrightarrow$" stands for material equivalence. In virtue of (1) we obtain,

$$2) u(h) = u\big(h\&(h \leftrightarrow m)\big)P(h \leftrightarrow m/h) + u\big(h\&\sim(h \leftrightarrow m)\big)P(\sim(h \leftrightarrow m)/h).$$

Since the truth-conditional identity of h and m guarantees that $h \leftrightarrow m$ is true necessarily, $P(h \leftrightarrow m) = 1$ and $P(\sim(h \leftrightarrow m)) = 0$. Accordingly,

$$3) u(h) = u(h\&(h \leftrightarrow m)).$$

Finally, since $[h \cdot (h \leftrightarrow m)] \leftrightarrow h \cdot m$, their outcomes must be identical. Therefore,

$$4) u(h) = u(h\&m).$$

By symmetry,

$$5) u(m) = u(h\&m)$$

and therefore,

$$6) u(h) = u(m).$$ [6]

[5] Here I approximately follow Bermudez 2009, pp. 91-92.

[6] This partition only works if one assumes traditional formulations of decision theory susceptible to Newcomb-type problems. The proposed advancement known as *causal decision theory* admits only partitions of what Lewis (1981b) calls

This proves that truth-conditionally identical actions must be ranked at the same level of preference by any utility function respecting the principle of expected utility. Accordingly, any rational agent in the light of Bayesian decision theory must be indifferent between two actions with the same truth-conditions. In sum, decision theory does not trade on representations of states of affairs but on states of affairs themselves and so cannot be intensional. If, as I have argued, psychological interpretation relies on decision theory, it cannot admit symbols as explanans these must be congruent with a non-intensional explanatory rubric. Worse, probability alone does not trade on representations either and so any attempt to incorporate degrees of belief and desire into psychological explanation must dispense with the alleged intensionality of the mental (e.g. Chisholm 1956).

IV. Schick's Challenge

Schick (1991) has argued that, as a matter of fact, people make different decisions based on different ways of conceiving (or, he calls them, *understandings* of) the possible outcomes of their deliberation. He provides an example taken from George Orwell's essay "Looking back at the Spanish Civil War" where he narrates his inability to shoot at a Fascist half-dressed and holding his trousers with both hands as he ran through an empty battlefield. Orwell explains that "[he] had come to shoot at 'Fascists'; but a man holding up his trousers isn't a 'Fascist,' he is visibly a fellow-creature, similar to yourself, and you don't feel like shooting at him" (Orwell 1957, p. 199). Schick concludes that Orwell's decision-making was sensitive to viewing or representing the man as a 'Fascist' or as a fellow-creature holding his trousers, despite being under no confusion as to the man's identity (Schick 1991, Ch. 3). If so, a descriptively adequate decision

"dependence hypothesis." However, since both formulations coincide in normal cases, and since the intensionality of intentional states manifests quite pervasively, I believe the result sufficient for my purposes.

theory cannot be extensional.

Another of Schick's examples presents a doctor who has to decide whether to provide care for her patient who has turned out to be a brutal murderer. The murderer, we may suppose, is seriously injured and her life is in jeopardy. According to Schick, the doctor might well make a different decision depending on whether she viewed the person as her patient or as a brutal murderer. If so, again, the way one views a person makes a difference to one's decision-making.

It is no coincidence that all the examples upon which Schick rests his case have an ethical dimension. Ethical reasoning epitomizes the volatile nature of outcome evaluation—a poorly understood part of decision theory. When making ethical decisions, we see ourselves as bound by multiple norms that apply or fail to apply to an action/outcome depending on nuances, which may turn our evaluations around 180 degrees. For example, in the famous Trolley problem our evaluation of the action/outcome flips depending on merely procedural conditions. The superposition of various such norms satisfactorily accounts for the difference in the decisions observed by Schick.

But the volatility of outcome evaluation is a general phenomenon; ethical reasoning just epitomizes it. To illustrate the point, let us assume that we value states of affairs and view these as sets of possible worlds (Lewis 1986c). We therefore value sets of possible worlds. An outcome is an individual possible world and, therefore, belongs to many states of affairs. When we settle on an intention to do something (e.g. to shoot at 'Fascists') we seek to bring about a state of affairs that we deem valuable. Call this state of affairs our goal. But in deliberating about the steps that will lead us to an outcome that belongs to our goal, we do not generally screen off all other considerations (Bratman 1987). Our intentions usually are parts of larger plans and when not, we seldom seek them at the cost of all our other goals and norms. As a consequence,

in deciding what courses of action will lead us to the achievement of our goal we must weigh the options not only in light of the value that we seek to realize with that goal but in light of our overall system of values. Only such a holistic evaluation of the outcomes will safeguard our general interest.

Consequently, since we value states of affairs, in evaluating an outcome we ought to recognize all the states of affairs to which the outcome belongs and toward which we are not indifferent. Each one of these will assign a value to the outcome. We must then balance off all these values in order to produce an input for the calculation of utility. Accordingly, even if I settled on an intention to X given that I value X highly, contingent conditions might determine that the outcome of X'ing presently, in addition to having the value I expected it to have, belongs to states of affairs that I view very negatively, thus deterring me from X'ing. This does not mean that I must give up my intention to X altogether. I may, for example, expect the conditions to change in a way that allows me to achieve my goal without incurring the previous costs. In sum, a given way of viewing an outcome may sway our decisions not because we base them on different symbols of the same situation, but rather because such a way of viewing the outcome may exemplify a contingent value that was not expected to result from the action when one settled on the intention to perform it.

Consider the case of the doctor faced with the decision to provide medical attention to a patient who has turned out to be a brutal murderer. On the one hand, the doctor has pledged to strive for the well-being of all her patients and, on this score, she deems the recovery of the individual resulting from her care a valuable outcome. On the other hand, she also believes that the world would be a better place without brutal murderers and deems, on this score, the recovery of the individual a negative outcome. Even if she has long settled on the intention to

serve her patients, in reaching a decision in this instance the doctor must balance off these and, perhaps, other considerations. But the fact that the doctor's deliberation is altered by the information that her patient is a brutal murderer does not imply that she can choose, as it were, how to view the individual. In fact, anyone who appreciates the doctor's dilemma will have to regard her decision as defective if made in view of the individual only as her patient to the exclusion of the fact that he is also a brutal murderer or vice versa.

Orwell's case seems a more plausible example of Schick's thesis because Orwell himself says that the individual holding his trousers as he runs on the battlefield is not a 'Fascist'—on account of this very description. However, given that Orwell knows very well that the individual is in fact a 'Fascist,' and that he knows that we know that he knows, etc. we must interpret him figuratively. There can be little doubt that he is communicating that the circumstances in which the individual appears bring to prominence one's identification with mankind, and eclipse any feelings of aversion, for they produce the compassion one should naturally feel for anyone in such an indecorous position. As a humanist, Orwell unsurprisingly attached a high value to humanity even during his days as a soldier.

Aside from this, Orwell's case is no different from the doctor's. Schick probably saw in the prominence of Orwell's feeling of identification a screen for any negative value he assigned to the individual in virtue of being a 'Fascist,' resulting in his decision not to shoot at him. As with the doctor, this construal of Orwell's deliberation does indeed favor Schick's thesis but at the cost of portraying Orwell as less than fully rational. If correct, Orwell's humanist sensibilities would have interfered with the incorporation into his deliberation of all the relevant information he possessed. Though it is not entirely unusual that people fall short of full rationality in this way, other things equal, an interpretation that avoids this consequence is to be preferred

(Davidson 1973b, Lewis 1974). Alternatively, we may view Orwell's recognition of the intrinsic value of mankind—because such recognition is essential to the possibility of organized society—as binding him to basic social values such as the intrinsic value of human life. This value requires that one does not take a human life unless to protect other human lives. Since a person holding his trousers as he runs can hardly inspire fear for anybody's life, it is plausible that shooting at him, 'Fascist' though he was, would have reflected indifference for the intrinsic value of human life. Conversely, even when Orwell did shoot at 'Fascists' he regarded his targets as valuable in virtue of their humanity but thought himself under no obligation to preserve their lives given that the lives of others, including his own, were on the line (surely, alongside other ideals dear to him such as equality and justice). But none of this requires that Orwell evaluate his target differently when he views him as a 'Fascist' and when he views him as "a man holding his trousers as he runs."

Generally, Schick's thesis is implausible because, unless one is under time pressure, it is irrational to ignore relevant information in making one's decisions. If decision theory is correct, we are value maximizers: our ultimate goal is to bring about the world that we most value among the options open. Our ultimate goal is not to bring about ways of viewing the world that we value. Even if we can view the world in a way that we like, the world might still be very much to our dislike. This is Oedipus' tragedy: he can view his mother as 'Jocasta' but he is all the same married to his mother. As a consequence, when assigning value to our options in deliberation we must give proper weight to all the relevant possible ways of viewing each option, hence to all the possible ways in which the option has positive and negative value, hence to all the relevant states of affairs to which it belongs. For once we have chosen an option and successfully executed the action, it is the option itself, not the ways in which we view it, that materializes.

V. Fodor's Solution to his Trilemma

I have presented Fodor's trilemma not only because it illustrates the perils of construing mental representations as symbols, but also because his formulation of the problem highlights its dependence on a particular set of cases such as that of Lois Lane or Oedipus. This stands in stark contrast with the generality with which Frege and subsequent philosophers have looked at the same type of problems. They have sought a theory of the mind that dissolves the anomaly; that is, a theory that renders the cases that cause the problems unexceptional and, in doing so, they have neglected important folk-psychological evidence. But let's not forget: if it were not for agents such as Lois Lane or Oedipus, there would be little problem with folk psychology being non-intensional.[7]

Having acknowledged this, Fodor's solution to his trilemma exploits, rather than dissolves, the anomaly. He begins with a characterization of the problematic cases in terms of his notion of epistemic equilibrium: an agent is in *epistemic equilibrium* iff the acquisition of new information would not change her decisions (1993, p. 42). He then proposes what he calls the principle of informational equilibrium which says that agents are normally in epistemic equilibrium with their environment and this explains their high success rate. Thus Fodor portrays the failure of agents such as Lois Lane as informational rather than rational: Lois Lane's problem can be cashed out in terms of false beliefs, namely, she falsely believes that Superman is a different person from Clark Kent. Fodor then proposes a special treatment for the problematic cases, namely, no treatment at all. He claims that all agents out of epistemic equilibrium are *ipso facto* beside the domain of application of the laws of psychology and so, one must conclude, psychologically incomprehensible.

[7] I assume that beliefs about non-existing objects can be dealt with independently. See chapters 5, 6 and 7 below.

Why should being out of epistemic equilibrium make the laws of psychology inapplicable to one? No independent argument is offered. One gets the impression that it is simply the need to avoid the trilemma that sustains this claim and that the idea of epistemic equilibrium is just an artifact for characterizing the problematic agents as unworthy of psychological consideration. But this will surely not do: Lois Lane may be in love with a flying man, but she is no madwoman—whatever neologisms Fodor uses to describe her. In other words, Fodor's idea of epistemic equilibrium does nothing to dispel the appearance that we understand Lois Lane, and so his proposal is no solution to his trilemma but rather the surreptitiously unmotivated negation of one of its horns. To make matters worse, Fodor's proposal is self-defeating, since by sidelining the problematic cases he is denying the intensionality of the mental and thereby cancelling the main motivation for computationalism, the thesis he is trying to salvage by this very maneuver.

But for all its flaws,[8] Fodor's general strategy of exploiting the anomalies has at least the virtue of originality in a field where philosophers have stubbornly been stumbling against the same wall for over a century. I claim that once we divorce it from Fodor's dubious agenda, the strategy can be turned into a real solution to our dilemma.

With some much needed refinements, of course: for one, Fodor's notion of epistemic equilibrium mischaracterizes the problematic cases and must be amended. It is true that the problematic agents, such as Lois Lane, are out of epistemic equilibrium, as defined by Fodor. But many agents out of epistemic equilibrium are not problematic in the required way. Consider L again—Fodor's favorite candidate for psychological law. Suppose that Peter wants to have a beer and believes (falsely) that if he looks in the bathtub he will get a beer. L seems to apply to

[8] For further criticism of Fodor's solution see Arjo 1996 and Aydede & Robbins 2001.

Peter perfectly well, as it yields a clear-cut, reasonable prediction: Peter will perform an action intended to result in his looking in the bathtub. And yet in this situation Peter is not in epistemic equilibrium, for acquiring the information that beers are not stored in the bathtub would have certainly changed his decision. The problem is that, quite generally, people relying on false beliefs are out of epistemic equilibrium while the cases we want to isolate are not cases of mere false believing.

The defect of the problematic agents is indeed related to false beliefs, but it is crucially related to beliefs about identities. For example, Lois Lane crucially relies on the false belief that Superman is not Clark Kent or Oedipus notoriously assumed that Jocasta was distinct from his mother. It is not obvious why false beliefs about identities should differ from other false beliefs in how they affect the psychological interpretation of their holders. But, as in previous chapters, decision theory provides a rubric to understand psychological explanation that gives satisfactory answers to deep questions about it.

The development of modal logic has taught us that our concept of possibility crucially depends on the basic framework of objects we recognize, whose identity and distinctness is necessary (e.g. Williamson 2001). In particular, it is impossible that an entity could have been two distinct entities. As a result, an agent's beliefs about identities determine what she views as possible in a foundational way. Whether or not what an agent views as possible really is possible, if decision theory is to captures her deliberation, it will have to assign subjective probabilities or credences to the possibilities as *she* conceives them. Accordingly, beliefs about identity frame an agent's deliberation structurally, in a way other beliefs do not.

Other false beliefs do not, in general, pose a problem for the application of decision theory to an agent's deliberation, for decision theory is neutral with respect to the relative

credences of all contingent conditions. A false belief about the possession of a given property by an entity is incorporated into the model of an agent's deliberation by simply assigning to the relevant state of affairs a higher or lower subjective probability (or credence) than we, interpreters, assign to it. But false beliefs about identity alter the very space of possibilities over which subjective probabilities are assigned and so no tweaking with the distribution of subjective probabilities will make a model framed in a different framework of objects apply to it. This is, in essence, the problem with Lois Lane and the like.

In the following chapters I will exploit the fact that the problematic agents always suffer from some identity confusions to develop an account of their psychological interpretation that dispenses with symbols and the intensionality of the mental. The explanans of psychological interpretation, I will argue, are relations between an agent and states of affairs defined by the role these states of affairs play in a decision-theoretic model of her deliberation. Moreover, the psychological interpretation of the problematic agents will turn out to be anomalous in ways that are rather superficial and, therefore, the proposal will not be in tension with the seeming unity of folk psychology.

Chapter 5

THE BAYESIAN BRAIN

Metaphysical scruples often lead philosophers to extremes. No doubt metaphysical adequacy is a virtue. But too often philosophers jump to a neat metaphysics without waiting for the verdict of empirical adequacy. A notorious example is David Lewis' modal realism (1986c). Lewis went from the conviction that the actual world could not be anything special in a sound theory of modality, to the postulation of infinitely many causally independent worlds. Lewis' conviction proved right, but the decade since Lewis' death has given us grounds to suspect a rather different conclusion.

We have learned that the brain is fundamentally a statistical machine (or as I will call it a Bayesian machine) grinding out distributions of probabilities and minimizing statistical error when carrying out even seemingly deterministic cognitive processes (Friston 2010). So it is true that our thoughts inhabit possibilities but it does not follow that in doing so they mirror an independent reality of parallel worlds. In other words, what goes on in the brain is so different from the phenomena it cognizes that there is no reason to expect that its fundamental reliance on possibilities bespeaks entities.

Lewis was for the most part alone in his endorsement of modal realism. But he was in the majority in viewing possibilities as mind-independent realities composed of entities of sorts. Lewis' conclusion has inspired incredulity and ridicule, but philosophers have not had much to say to address the concerns that inspired it. To use Lewis' metaphor, all of us enjoy, sometimes abashedly, the theoretical benefits of possibilities without looking at the metaphysical bill. The bill, says Lewis, is commitment to possibilities as mind-independent realities on a par with actuality.

Here I will argue that Lewis' modal realism follows from an uncritical acceptance of the

manifest image of the world as a mind-independent reality. If actuality itself—as it appears to us with its objects and properties—is partly an epistemic construct then it can be on a par with possibilities without their being mind-independent realities. In other words, instead of turning possibilities outside of the brain to accompany actuality, we turn actuality inside and achieve the same result. The price we have to pay is, of course, the old Kantian separation between the manifest image of the world and the world in itself: we must accept that the manifest image of the world may be determined by how our brains work as much as by the world in itself. I do not claim that this is a bargain but that it is the best price we will get. At a minimum, it is surely cheaper than modal realism.

In so arguing, I will be putting empirical evidence ahead of metaphysics rather than the other way around. In two ways: 1) I will argue that purely metaphysical reasons for the mind-independence of possibilities are weak or, at best, inconclusive; and 2) I will show that a host of recent empirical findings in neuropsychology and behavioral economics strongly suggest that objects and possibilities are epistemic categories.

I. The Epistemic View of Possibilities

Both in time and space, every human being is a speck, grossly overmatched in its challenge to understand, predict, and ultimately control the world that surrounds her. If, instead, we were omnipresent and omniscient in the way we sometimes think of God, we would not need possibilities for we would not need to predict the world. In fact, even finite as we are, boundless knowledge and computational resources to process it (much like Laplace's demon) would make possibilities otiose—provided, of course, that the world is deterministic.[1]

[1] If, as Quantum Mechanics suggests, the world is intrinsically indeterministic, possibilities would be unavoidable for a physically finite cognizer. But, again, if the indeterminacy is circumscribed to the microphysical, even in such a world, a Laplacean demon with similar interests as most of us might find possibilities superfluous.

With or without determinism, epistemically limited beings must cope with massive, inherent ignorance of the world. Ignorance leads to uncertainty and possibilities provide an effective way of managing uncertainty. Agents have foresight, that is, they are capable of somehow modeling parts of world, and thereby anticipate its evolution. Perhaps we can think of the principles of logic as comprising the basic framework for these models and whatever evidence we take as given as their inputs. But inherent ignorance manifests in the fact that these models are partial, both in that they can only represent a small part of the world[2] and in that, even within this small part, they rarely incorporate all the factors of its evolution. The latter fact is the result of our inability either to collect all the relevant evidence or to process it. But either way, their partiality makes these models underdetermine their outcomes. In other words, several evolutions of the models are logically compatible with the inputs. We have no choice but to deal with *ways in which the world could be* or, simply, *possibilities*. As Descartes demonstrated, uncertainty is pervasive in that any piece of evidence that we could use as input to a model of the world must also be the result of managing further possibilities all the way down to sensory data. It is thus plausible that agents are, intrinsically, managers of possibilities.

Interpretations of Quantum Mechanics positing irreducible indeterminacy need not be antagonistic to this view. Some microphysical events may well be probabilistic in nature but it does not necessarily follow that possibilities exist independently of us agents and, in particular, independently of our uncertainty. The world's indeterminacy can also be fully accounted for in terms of predictability. That some microphysical events are probabilistic in nature could simply mean (consistent with Bell's theorem) that there is no further local factors of the phenomenon

[2] I use "small" not in the sense of size but in the sense of containing partial information. For example, we can imagine what will happen with a certain galaxy but our judgment would be based on a small amount of information about that galaxy.

that we could use to predict its outcome.[3] Irreducible unpredictability thus suffices to account for any irreducible indeterminacy. Such an explanation, also dependent on the presence of a cognizer, would have the benefit of metaphysical frugality over its competitor.

In contrast, though Descartes' skeptical scenario produces no conviction that knowledge is unattainable, it does conclusively show that uncertainty is invariant among finite agents. We all can be certain of the same, namely, almost nothing. Strictly, the only propositions we can be certain about are those whose truth is independent of the world (e.g. tautologies or "if I think then I exist"). In other words, one should not give a subjective probability of 1 or 0 to a proposition that is open to reconsideration (Lewis 1981). Aside from tautologies, circumstances of reconsideration are not hard to conceive for any proposition. Circumstances in which I would reconsider my sensory information can no doubt be outlandish, as Descartes' scenario illustrates, and this affords them a degree of confidence close to 0. But they are conceivable. This connects uncertainty with logical consistency: possibilities are, in their most fundamental sense, logically consistent states of affairs.

In practice, oftentimes we don't even know how far our uncertainty ranges. For example, our conceptual structure usually is blurry; our concepts usually lack sharp boundaries. Lack of conceptual refinement may lead to equivocation or error in assessing uncertainty. It may seem to somebody that there is no inconsistency in a creature exactly like me in its behavioral dispositions but without phenomenal consciousness (Nagel 1974). But this may well be due to her poor understanding of the mechanisms that enable the sophisticated behavior exhibited by humans. Or, less controversially, the layperson may see no inconsistency in there being a highest prime number yet this is due to her lack of understanding of mathematics. In practice, our judgments about consistency are fallible and do not always track possibility. But this does not

[3] I am referring to the de Broglie-Bohm interpretation of Quantum Mechanics.

show that the notion of possibility cannot originate in our sense of logical consistency. It may simply show that a rigorous, interpersonal treatment of possibility requires abstracting from particularities of human beings such as their conceptual refinement and computational capabilities.

Why then have philosophers eschewed an epistemic account of possibilities? The reasons are miscellaneous.

II. Russell and Kripke's Influence

One reason has been Russell's influence. Merely possible objects have enjoyed a bad reputation since Russell dismissed them from a "robust sense of reality." Moreover, Russell claimed that his analysis of names as definite descriptions shows how to dispense with them for analytical purposes. We need not analyze "Sherlock Holmes did not exist, but he might have" as a proposition containing the unactualized possible object Sherlock Holmes; we analyze it as "There wasn't a unique consulting detective who lived in Baker St., but there might have been a unique consulting detective who lived in Baker St." Thus did Russell reject Meinongianism (1905a). Russell's influence on Quine is evident in Quine's mid-20th-century discussion of "pegasizing" in 'On What There Is' (1948), which in turn persuaded many philosophers of that generation and the next to downplay the metaphysics of possibilities. Of course, Russell's analysis has been recognized to face grave problems since Kripke's *Naming and Necessity* (1980).

But, perhaps more importantly, even when not explicitly present in discussions of reference, Russell's view of the mind lingered in the minds of philosophers. Russell viewed the structure of the mind as mirroring the structure of external reality. This was the basis for his elaborate theory of logically proper names and reference by acquaintance in his *Philosophy of*

Logical Atomism (1956). But such a view yields a rather simplistic picture of the mind's fundamental cycle, namely: it takes information from the senses, constructs a straightforward representation of the external world, and then grounds behavioral decisions upon that representation. This fundamental cycle, with no clear role for possibilities, has loomed large in philosophy of mind, and has been taken for granted by most computational theories of mind. However, the more we understand the brain the less plausible this view has become.

Secondly, there is a widespread tacit assumption that possibilities pertain to metaphysics and so that epistemic make-up is illegitimate grounds to explain them. Here not Russell but Kripke is the originator. In *Naming and Necessity*, Kripke famously argued that "necessary" works as a metaphysical operator in contrast with "a priori" which works as an epistemic operator (1980). Given that a formalization of the logic of possibilities requires only the addition of the "necessary" operator to a non-modal language, Kripke's distinction led to the association of possibilities with metaphysics. Kripke's distinction is correct, in my view, but easily misinterpreted. The culprit is the alleged dichotomy between the metaphysical and the epistemic.

Though widely taken for granted, there is no clear-cut dichotomy here but an illustrative contrast. Philosophers use the terms "metaphysical" and "epistemic" to capture various contrasts between degrees of involvement of agents' cognitive make-up in the truth-conditions of certain statements. Putnam (1983, 1987) has altogether rejected the dichotomy on the grounds that all statements are, at bottom, dependent on our conceptual scheme and, thus, on our cognitive make-up. This has earned him the reputation of relativist. But we need not go that far to clarify Kripke's remarks. It suffices with the harmless distinction between a statement depending on the particular cognitive make-up of an individual and a statement depending on cognitive make-up in the general sense of that which is common to all individuals.

Kripke's goal in the said passage was to establish that whether a proposition is necessary does not depend on a particular agent's knowledge, in stark contrast with whether a proposition is (known) a priori. This was important to give metaphysical possibility and necessity sufficient stability. But this observation is compatible with the view that whether a proposition is necessary depends on the kind of cognitive architecture shared by all humans (perhaps all agents). Of course, we can't blame Kripke for appealing to the dichotomy of the metaphysical versus the epistemic in articulating his point, for how else could he better put the point? However, his argument shows that necessity is not an *epistemic* notion, when this is construed in the narrow sense of involving the particular cognitive make-up of individual agents. The view that necessity is an *epistemic* notion, when this is construed in the sense of involving fundamental aspects of agents' cognitive architecture, is untouched by Kripke's point. Possibilities, when construed in this way, remain metaphysical categories, in Kripke's sense. But, in the sense that matters here, they arise from the brain meeting the world all the same.

My claim is not that possibilities arise from any commonality whatever of our cognitive make-up. In particular, from the fact everybody believes that Aristotle was someone else it does not follow that it is possible that Aristotle had been someone else (Kripke 1980). Rather, possibilities depend on the cognitive architecture common to all agents, that is, the kind of machines brains are. In contrast, intentional content, whether or not shared by all agents, makes no difference to possibilities. In fact, Kripke's argument against interpreting possibility as an epistemic notion is explicitly addressed to the content of an individual's cognitive make-up. It is *what* someone knows, not the kinds of processes and relation to the world that sustain knowledge, that Kripke views as inessential to necessity.

In developing a logic of knowledge, Hintikka 1962 defines a notion of possibility that

does depend on a particular agent's knowledge. (Needless to say, this notion of possibility runs counter to Kripke's argument. Yet this does not undermine Hintikka's efforts, for Kripke's point was only about what he took to be the fundamental meaning of modal notions.) Obviously, it would be an error to identify the notion of possibility I have in mind with Hintikka's notion. For, unlike him, I do not claim that the possible is what is compatible with what one knows (Hintikka 1962). In fact, while such a claim makes sense for defining a logic of individual knowledge, it would be at odds with a definition of possibilities based on our shared cognitive architecture because what one knows is particular to one and certainly not part of one's basic cognitive architecture.

III. The Correspondence Theory of Truth

Finally, an account of possibilities in terms of uncertainty must deal with the deeply ingrained, if perennially sketchy, correspondence theory of truth. According to my proposal, the truth of a subjunctive statement irreducibly depends on our epistemic make-up and, hence, there is no fact whose structure is isomorphic to the lexical structure of the statement which makes it true. This refutes the correspondence theory of truth or, at least, restricts its applicability. But the conflict has little to do with the reliance on epistemic make-up for any account of the metaphysics of modality, except possibilism and ersatz actualism (to use Lewis' epithet), will *ipso facto* contradict the correspondence theory of truth, as defined. In fact, a case could be made that even ersatz actualism contradicts the correspondence theory of truth because, on any such account, subjunctive statements would be true not in virtue of the intended references of its lexical terms but in virtue of some substitutes for them (e.g. abstract essences in Plantinga 1976).[4]

[4] Still, the conflict could be adverted by tinkering with the notion of correspondence involved in the correspondence theory of truth.

Either way, I do not think that the correspondence theory of truth should be considered an insurmountable obstacle. For, although motivated by a very plausible intuition, it is an intuition sparked specifically by declarative statements. What's worse, the correspondence theory of truth has serious troubles accommodating even these. In effect, logical connectives find no obvious metaphysical category corresponding to them nor are the true statements they help to form isomorphic to the intuitive facts that make them true. For example, disjunctive statements are made true by the correspondence of any of its disjuncts with a fact, or conditional statement may be true in virtue of the correspondence of the negation of the antecedent with a fact. This has led some to restrict the theory to atomic propositions (Russell 1914). But this is the least of the problems for the correspondence theory of truth. Worse is the absence of a metaphysically sensible fact corresponding with true negative existential statements. What fact could possibly correspond with the statement that there is no golden mountain (e.g. Williamson 2000)? In light of its problems, we should be careful with the word "theory" we use to refer to the correspondence theory of truth, for it suggests a full-blown, general account of truth. In reality, it is at best a vague insight into the nature of truth.

But aside from the problems the correspondence theory has with declarative sentences, there is no analogous metaphysical intuition sparked by subjunctive statements. The truth of the statement that I could have been the president of Zambia does not obviously rest on a fact composed of me, Zambia and the binary relation *is-the-president-of*. If such a fact exists, it surely corresponds with the statement that I am not the president of Zambia. But to insist that there is a fact that makes it true that I could have been the president of Zambia would require distinguishing it from the fact that corresponds to my actually not being the president of Zambia, and therefore would require the postulation of intrinsic properties associated with tense. In short,

it would require the postulation of intrinsic modal properties. Not only would this grossly multiply intrinsic properties, but there is nothing intuitive about such a metaphysical hypothesis. Hence, the adoption of the correspondence theory of truth for modal statements is quite a leap from its initial motivation.

There are a number of areas of human discourse forcing the correspondence theory of truth upon which leads to artificial conundrums. Consider moral statements. However controversial metaethics remains, virtually everybody agrees that the truth of moral statements does not depend on facts isomorphic to their lexical structure.[5] At the very least, it depends on facts having to do with the totality of human beings, the communities in which they live and the kind of creatures they are. Forcing correspondence upon ethics would entail a form of moral naturalism according to which moral predicates correspond to moral properties intrinsic in actions regardless of the existence of any agents other than the performer. This would drastically undercut the theoretical resources available for understanding moral problems. For one thing, it would ban agent neutrality as a relevant moral consideration. It is therefore implicit in all major ethical frameworks, including Utilitarianism, Kantianism and Rawlsianism, that the correspondence theory of truth does not govern moral discourse.

Given the metaphysical commitments of possibilism (e.g. Lewis 1986c), abdicating the correspondence theory of truth for modal discourse seems a minor concession. In fact, I claim that modal truth is similar to moral truth in that, without in general depending on any particular individual, it depends on the kind of creatures we are. This commonality makes both moral and modal discourse unfit for the correspondence theory of truth. But neither contradicts the truth-maker principle, according to which for every contingent truth there is something that makes it

[5] I include in "almost everybody" those who deny that moral statements have truth-values (viz. non-cognitivists; e.g. Ayer 1952) and those who construe truth as fact-independent (viz. minimalists about truth; Horwich 1990)

true (Armstrong 1997, 2004). To be sure, there are facts about the world that make moral and modal truths true, but these facts are not isomorphic to the lexical structure of such truths.

IV. Truthmakers and Mental Models

There remains the challenge of providing a plausible account of the truth-conditions of modal statements vis-à-vis a metaphysically frugal treatment of non-actual entities. Menzel (1990, 1991) provide just such an account. Take the possible-world interpretation of modal discourse at face value: each possible world is a model of how things could have been, which includes a set of individuals and a set of properties. Menzel lays down conditions under which a set of models of this kind will represent possibilities. These conditions include, obviously enough, that exactly one model matches reality seamlessly both in entities and properties and, less obviously, that the models overlap on individuals so as to guarantee transworld identity. Roughly, the idea is that there is exactly one model that embeds seamlessly in each way the world could have been and that, given any way the world is, close variations are captured by models which differ from the actual model (hereafter @-model) only in those aspects in which things would have been different.

This yields homophonic truth-conditions for modal statements. For example, "there could have existed more objects than there actually exist" is true iff there is a model, M, that contains all the individuals of the @-model plus, at least, one more, call it e. This will be the case iff reality could have been such that M would have fit seamlessly in it and, since M contains the individuals of the @-model, e (and perhaps others) would have corresponded to something that does not actually exist; hence iff there could have existed more objects than there actually exist.

Menzel's account has come under attack on the grounds that these truth-conditions do not reveal what grounds modal truth (Linsky & Zalta 1994). Menzel (1991) has responded that an

account of the semantics of modal discourse is under no obligation to reveal the ultimate metaphysical structure of reality underlying our use the term "could", any more than Tarski's scheme "'grass is green' is true iff grass is green" should reveal the ultimate metaphysical structure underlying our use of "green."

While I find Menzel's response satisfactory, the truth-maker principle guarantees that there must be grounds in reality which make modal truths true. I claim that those grounds are our fundamental epistemic make-up. For example, what grounds that there could have existed more objects than there actually exist is the fact that the ways in which reality could be, compatible with what one can be certain about, are boundless: whichever of these ways reality is, there are other ways that contain those objects and more. In other words, Menzel's account is consistent with an interpretation of subjunctive expressions as grounded in our fundamental epistemic make-up and therein expressing facts about agent-independent (though not mind-independent), irreducible uncertainty.

Although Menzel rightly places the semantic import of possibilities in the overall structure of the models, something must be said about what these models are. Menzel suggests arbitrary sets for the role of objects and worlds (1990).[6] The reason probably is that he considers sets metaphysically respectable entities—or, at least, more so than abstract entities and other Platonic alternatives (Linsky & Zalta 1994). Yet even this is granting too much to the metaphysical view of possibilities. Once their epistemic nature is acknowledged, the identification of Menzel's models with intentional events suggests itself. Our ability to imagine possibilities may be elusive and even mysterious, but it is incontestable. We can thus use the formalization of modal logic as a guide to understanding the place of possibilities in the mind

[6]Menzel follows standard formalizations of modal logic in defining properties and relations as sets of objects associated with possible worlds.

and exchange the metaphysical questions about possibilities for the question about the intentionality of the mental (Brentano 1874).

Let us associate possibilities with intentional contents and, in keeping with model theory, let us call these *mental models*. Psychologists have similarly defined and successfully used mental models to explain much high-level cognitive activity (e.g. inferential reasoning and language comprehension) while remaining mostly neutral with respect to the philosophical questions they raise (Johnson-Laird 1983). For this reason, borrowing mental models to characterize the types of intentional events that we ordinarily would call "conceiving a possibility" should not be too philosophically tendentious. Plus, when we conceive possibilities we do not have in mind full worlds, specific in every detail, but partial situations (Barwise & Perry 1983). This sits well with the way mental models are defined in the psychological literature (Johnson-Laird 1983).

Let us assume that associated with an agent is a set of intentional objects and a set of relations (including unary relations or properties) and that these are combined to form mental models. Following Menzel's models, let us assume that intentional objects are not inextricably wedded to any properties or relations.[7] In a sense, intentional objects are empty for no property is essential to them. But since they are the building blocks of mental models, they constitute the lower limit of variation: they are invariable across models. Consequently, the only property they have essentially, namely, their identity, they have it necessarily: each object is identical to itself and distinct from every other object. In short, they are mental versions of bare possibilia (Williamson 1998) and so I will sometimes call them *mental possibilia* to emphasize this fact.

However, the sense of possibility captured by mental models is not exactly the same as

[7]Though they might exhibit prototypic effects when they are summoned to partake in mental models in which prototypic properties make no difference (e.g. Rosch 1978).

the formal one developed in modal logic since mental models represent, at best, a finite array of circumstances whereas possible worlds, as standardly understood in the literature, can include infinitely many states of affairs. In fact, on the intuitive interpretation of modal language the domain of objects is typically thought to be infinite, which disqualifies mental models as possible worlds. Intuitively, mental models should be thought of as partial models of possible worlds.[8] Following Lewis 1986c, we may think of a mental model as standing for sets of possible worlds, namely, the worlds consisting in logically consistent variation of the states of affairs left indeterminate by the mental model.

But despite their limitations, mental models are all we needs to manage possibilities. First, the finitude of the domain of intentional objects is not an insurmountable obstacle because intentional objects do not exhaust the resources for constructing mental models. Rather, agents can introduce auxiliary objects at will without a specific identity. In fact, it is independently plausible that unidentified objects play a crucial role in many uses of mental models as, for example, the interpretation of pronouns anaphoric on definite descriptions (Garnham 2001). Since intentional objects do not have any property essentially except their identity, no possible object is beyond incorporation into an agent's mental models, if only as an unidentified object. If so, we can assume that agents have an unlimited supply of auxiliary objects and, accordingly, interpret the domain of mental possibilia as the domain of all conceivable objects.

Consider now a sentence composed only of constants and predicates, that is, without quantification or nested modalities. The proposition expressed by such a sentence is possible iff there is a possible world at which it is true. Now if there is a mental model that satisfies the sentence it follows that there is a possible world at which the sentence is true: in effect, any of

[8]Or perhaps as *situations* (Barwise & Perry 1983) only without their mind-independence.

the possible worlds logically consistent with the model will do.[9] Conversely, because a sentence in the modal language is the result of the finite recursive application of formation rules, it cannot contain more than a finite number of constants and predicates. Consequently, if a sentence composed of only constants is true at some possible world, there must be a finite part of that world in virtue of which the sentence is true, and so there must be a model that makes the sentence true. In sum, the proposition expressed by a sentence composed of only constants is possible iff there is a mental model of it.

The real challenge for mental models is infinitude as comprised in quantified sentences. Being partial models of possible worlds, mental models cannot straightforwardly embody quantified sentences because the propositions these express often involve infinitely many states of affairs. However, mental models need not represent quantified sentences to serve as semantic ground for them. Suffice it that the system of mental models allows us to assess the truth value of all quantified sentences.

Assuming the locality of logical inconsistencies (that is, that all logical inconsistencies are somehow detectable with partial models[10]) a standard fixed-domain interpretation of quantifiers can be given in a system of mental models along these lines. The standard interpretation in terms of possible worlds is that $\Diamond \forall x \alpha$ is true iff α is true of every possibilia at some possible world.[11] A possible world is intuitively associated with a logically consistent way

[9] That the set of possible worlds logically consistent with the mental model is non-empty follows from defining possible worlds in a sufficiently ecumenical manner. I do not, for example, assume that possible worlds satisfy a maximally consistent set of sentences, as has often been stipulated. Perhaps a possible world results from filling out all the details required by the mental model and no more, leaving scores of sentences without a true value. But even if possible worlds were defined in terms of maximally consistent sets of sentences, I believe that a possible world logically consistent with the mental model could be constructed with, roughly, Henkin's method (1949; see also Bayart 1959, Cresswell 1967 for applications to modal systems).

[10] I mean to be neutral with respect to the existence of logically inconsistent mental models (Johnson-Laird, Byrne & Shaeken 1992; Johnson-Laird, Girotto & Legrenzi 2004). Shouldn't there be any, one must interpret my references to logically inconsistent models in the text as an abbreviation for impossible models or, perhaps, for model-formation processes that cannot be carried out.

[11] A fixed-domain interpretation validates Barcan formula and its converse and, therefore, quantifiers and diamonds commute. Since also we can get the particular quantifier from the general quantifier plus negation and the box from diamond plus negation, I assume this case suffices for the present purposes.

in which everything beyond a mental model could be. Hence, $\Diamond \forall x \alpha$ is true iff there is a mental model logically consistent with α being true of every conceivable possibilia (whether or not it belongs to the model and whether or not it exists according to the model) iff it is not the case that some conceivable possibilia are such that α being true of them is logically inconsistent with every mental model. Since the empty model is one such model, this comes down to the requirement that there are no conceivable possibilia such that α being true of them would be logically inconsistent. In virtue of the locality of logical inconsistency, this is the case iff there is no logically inconsistent model consisting only of α being true of its members.

I believe that these assertability conditions, reminiscent of Popper's fasificationism about science, are psychologically plausible. When thinking about whether some generally quantified statement is possible, we usually proceed by the method of counterexample, that is, we try to concoct a finite situation that reveals the logical inconsistency of the condition dominated by the general quantifier. For example, is it possible that all men in a town are such that they are shaved by the barber if, and only if, they do not shave themselves? Russell answered this question in the negative observing that if all men in the town satisfied the condition, it would follow that the barber shaves himself iff he does not shave himself (1914). In other words, to answer the question whether a generally quantified statement is possible, it seems natural to show that a model of the condition dominated by the general quantifier cannot be constructed.

A variable domain interpretation of quantifiers (e.g. Kripke 1963) would require, first, stipulating that mental models symbolize how things are with the objects that exist in a possible world. Which world? Anyone logically consistent with the model. This will complicate the representation of reasoning involving the existent and the non-existent simultaneously (e.g. Pascal's Wager). But the truth-conditions of quantified statements in terms of mental models

would not be significantly different from the fixed-domain case except for a relativization of the conditions to those mental possibilia that exist at a world. However, the point of developing a mental model account of possibilities is precisely to soothe the metaphysical fears that motivate the variable-domain interpretation of quantifiers in modal language.

The variable-domain interpretation was custom-made to invalidate both the Barcan formula and the converse Barcan formula (Kripke 1963). The problem with these formulas is that they entail $\forall x \Box \exists y\, x = y$, which is interpreted as stating that everything exists necessarily (Prior 1957). But once we dissociate the particular quantifier from existence, nothing stands on the way of the Barcan formulas. In fact, the purpose of providing an interpretation of modal language based on mental models (the purpose of Menzel's models, for that matter) is to ease the metaphysical fear of viewing the domain of quantification as containing objects that do not exist. The system of mental models proposed here starts from a framework of mental possibilia invariable across models, hence necessary, and, hence, $\forall x \Box \exists y\, x = y$ should intuitively be true in the non-existential interpretation (Williamson 1998). Beside this metaphysical fear, a case can be made that the Barcan formula and its converse are intuitively plausible (Williamson 1998), the system they yield natural and parsimonious, while Kripke's variable-domain system is saddled with conundrums (Williamson 2000; Cresswell 1991).

V. The Enactment of Possibilities

Having argued for the plausibility of a mental-model account of modal truth, in the last two sections I would like to explore the question about the nature of mental models. This question is closely related to the question of how the Bayesian principles are implemented by the brain. Details aside, there seem to be two fundamentally different options. First, mental models may be neural symbols that undergo transformations in cognitive processes. This would be in line with a

computational view of the mind and tantamount to the idea that possibilities are explicitly represented in the physical realization (viz. hardware) of the machine (see chapter 4). Second, mental models may be part of the algorithmic realization (software) of the machine, that is, they may arise as high-level, functional descriptions of the physical processes taking place in the brain. This would be in line with the recent trend in cognitive science known as situated cognition (for reasons that will emerge) and tantamount to the idea that possibilities are implicitly represented in the brain's workings.

In favor of the first option is the fact that, on the Bayesian hypothesis, possibilities are the gist of cognition and so essential to the way agents operate in the world. As a consequence, mental models should be real entities participating in the causal fabric of reality. In short, possibilities should be causally efficacious and neural events certainly have the required causal powers. In favor of the second option are our intuitions. I am prepared to take them seriously.

Intuitions have it that possibilities often are imaginable, that they are *what* we imagine when we conceive them. If so, mental models must be the contents of our inner experiences: they must be what our imaginations are about. Whatever mental models are, we certainly do not in general imagine neural events when we imagine possibilities. Of course, it may be that these inner experiences are realized in the brain in a way that mirrors their contents and so that possibilities are also explicitly represented (Barsalou 1999). But, first, this is a substantive thesis for which decades of spectacular advances in neurophysiology have produced little support. More importantly, if we are to honor our intuitions, this would make no difference, for mental models should still be associated with the contents themselves that are imagined, not with the symbols that would realize those contents. When I imagine Napoleon and Plato shaking hands, I do not imagine representations of Napoleon and Plato shaking hands. It is Napoleon and Plato

themselves who I imagine shaking hands. This is no terminological point. It is crucial for a Bayesian machine that possibilities are not distinguishable by the role they play in its cognitive architecture: what is actual and what is merely possible must be a contingent matter (Kripke 1980, Lewis 1970b). Yet clearly we experience the actual world directly, not just indirectly through representations (Searle 1983, Ch. 2). So must we experience all possibilities directly. And this is exactly the way we speak about the world and its possibilities. Forcing the computational mold upon the Bayesian brain would thus require, at a minimum, claiming that our commonsense view of the world and its possibilities is inadvertently metaphoric.

The issue of the causal efficacy of possibilities is greatly exaggerated. In effect, it does not follow that mental states about possibilities are causally inefficacious from the fact that mental models arise from high-level, algorithmic characterization of brain states and processes. For, however brain states are characterized, their causal powers remain unaltered. Nor does it follow that cognition must dispense with possibilities. In a sense that reflects the divide between the mind and the brain, when we ordinarily talk about cognition we are talking about *what* the brain does, not *how* it does it. "Cognition" is a folk notion coined much before we knew about neurons which has recently been recruited for scientific service. Yet even in scientific contexts it retains its core folk meaning conveying algorithmic descriptions of brain processes. In fact, it is frequently used to draw a contrast with neural descriptions of these processes. If so, placing mental models at the level of algorithm is placing them where they must be to shape cognition.

Never mind if brain states are not causally efficacious in virtue of the possibilities they realize (Kim 1984, Block 1990). Suffice it to explain this appearance that sometimes the contents associated with brain states reflect the causal relations between them at the neural level. Consider, for example, the task of figuring out whether a certain object can be rotated into a certain

position. There is strong evidence for the thesis that people normally perform this task by imagining the object (Kosslyn 1980, 1994). In fact, when one is faced with such a task, one need not be particularly introspective to notice that one is conjuring up the experience of manipulating the object in Euclidean space and so, inter alia, constructing a dynamic mental model of the transition from the initial position to the final. The mental model involved is thus the simulation of the visual-motor experience of rotating such an object in Euclidean space. But even assuming that there are no explicit representations of the objects in the brain, the brain states forming the sequence implementing the simulation may be associated by their contents if contents crucially capture patterns of visual-motor contingencies that are activated at the neural level (Noë & O'Regan 2001, Noë 2004). If so, the neural realization of one (together with the realization of the appropriate intention) would enable the neural realization of the following one in the sequence.

Other times it seems perfectly natural that mental states' causal powers are rather arbitrary in light of their contents. Consider the case of reinforcement learning (Sutton & Barto 1998). If we have been sufficiently rewarded with dopamine secretion for a certain action in a certain context, the very anticipation of the action releases dopamine, increasing the likelihood that the action will ensue (Schultz 1998). It does not matter much what action it is, provided it is not immediately harmful. Nor does it matter, therefore, that we don't have a clue as to why the action is rewarded or why the anticipation feels good. In other words, it does not matter that we do not see the content of the mental state of anticipation as the reason to perform the action. And if the mental state's content does not in any way explain the occurrence of the action, we cannot see it as causing it.

As these examples suggest, I am thinking of imagination as a kind of brain simulation.

Assuming the token-identity of experiences and brain states (e.g. Lewis 1966), by simulating brain states, one thereby simulates experiences. Since experiences have contents given by objects and properties, the simulation of experiences allows us to visualize and evaluate scenarios. These scenarios correspond to what I have called mental models and embody possibilities in our cognitive processes. The imagination of a possibility consists, therefore, not so much in its (explicit) representation in the brain as in its enactment by it.

The idea that advanced organisms represent the world by simulating experiences of it has been conjectured in recent years by several advocates of the so-called situated cognition movement. What has driven these conjectures is the mounting evidence of sensory-motor integration in the brain. Areas known to specialize in motor control have been shown to have sensory properties and areas known to specialize in sensory processing have been shown to play a major role in motor control (Gallese & Lakoff 2005). This integration does violence to the traditional computational paradigm according to which the brain has a central processing area which is fed by the sensory module and feeds the motor module, and where the world is represented and modeled by non-sensory, arbitrary symbols (Fodor 1983). Instead, there is ample empirical evidence that the brain represents the space, actions and various common objects by simulating sensory-motor experiences as of those things (Gallese 2005).

This conjecture is also congenial to evolutionary speculation. It is plausible that the brain evolved initially as an organ endowed with a predetermined repertoire of reactions to some statistically crucial environmental conditions. Such a primitive function would surely be implemented without the participation of anything like what we call foresight. Accordingly, it seems unlikely that this primitive brain would have relied on explicit neural representations of the external world to encode its reactions. But either way, this repertoire of reactions could have

been refined through evolution, in tandem with an increase in brain mass and capacity, to allow for sensitivity to more and more nuanced states of the environment (Parker 2004). This refinement continued up to the point when the availability of brain capacity could have offered the opportunity to model the unfolding of environmental conditions. Foresight was thus born. But how could the brain model the unfolding of events in its environment? Since the brain does not have direct access to the environment, let alone to the future, but only to its own states, and since there already was a rich repertoire of reactions to environmental conditions, these may have been recruited to serve as implicit representations of the environment and its unfoldings. In other words, the brain somehow may have evolved the ability to use itself as a model of the world; to survey (with the body off-line) sequences of brain states and responses to various environmental conditions in order to evaluate them in broadly Bayesian style before settling on one.[12]

Since no representation of the world (a fortiori, of the environment) can contain inconsistencies,[13] on the assumption that the brain evolved to implicitly represent the environment, it is plausible that by construction of the brain, experiences always are consistent. One cannot, for example, imagine a square circle or an object that is both fully white and fully black. If so, experience provides a basic test for logical consistency.[14] Because I have been assuming that what is logically consistent is possible, it follows that what is experienceable is possible. Here "experienceable" must be understood in the broad sense, however, as not relativized to one's position. Understanding it otherwise would make a fool of the proposal for it would trivially imply the impossibility of Descartes' skeptical scenario: it is not experienceable that one is being deceived by a demon while being deceived in the way envisioned by Descartes.

[12] The self-use of the brain has been conclusively established in recent years (Kanheman 2011).

[13] Cf. Lewis' Rule of Actuality: "nothing false can be properly presupposed" …in one has knowledge (1996b, p. 554).

[14] I say "basic" because experience is subject to constraints of abstraction and computational capacities. For example, a 17-dimensional space is not experienceable yet, in an important sense, possible (Lewis 1970b).

The non-relative sense is that of being experienceable from some position or other, not necessarily mine. In this sense, the possibility of Descartes' skeptical scenario rests upon a simple inference: first, I can imagine witnessing any agent being deceived in the Cartesian way; hence it is possible for any agent to be deceived in that way. Second, I am an agent; therefore, it is possible that I be the one deceived. (Hence, agents who lack self-awareness should be unable to appreciate Descartes' point.)

Unthinkable only decades ago, the Bayesian brain is spectacularly illustrated by current theories of perception. Perception used to be thought of as the antipode of modality and hypothesis testing. It used to be thought of as modular and passive: perceptual systems receiving the sensory input from the world, processing it and sending to the abstract centers where a representation of the stimuli was formed. However, mounting evidence shows that this is far from the truth. The brain is Bayesian even in perception: not a passive receptor of information about the world but proactive in formulating predictions and testing hypothesis driven by minimization of statistical error (Kveraga et all 2007, Friston 2005). These predictions are generated in higher areas of the brain, that is, areas where relatively abstract processing takes place involving, among others, the integration of information from the various sensory systems. The predictions are then passed down to the specific sensory system receiving the sensory input to modulate its interpretation. For instance, the prediction can sensitize, or even bias, the sensory system to certain aspects of the stimulus to confirm or disconfirm the prediction. When the prediction is disconfirmed beyond a threshold of statistical error, another prediction is issued until a fit is found. In short, by resort to previous experiences and stored knowledge, the brain greatly narrows down the possible interpretations of the sensory stimuli facilitating recognition and processing.

Not only does this show that perception involves the contemplation of possibilities, much in the way Bayesian epistemologists have been fancying for a while, but it also provides further evidence for the claim that mental models lie in experience rather than in the underlying neural structures. In effect, the way in which the various systems involved in perception communicate is by synchronization of neural activity (Varela et all 2001, Friston 2005). This requires that the information comprised in the prediction generated in the higher areas and the information comprised by the sensory input in the lower areas is encoded in the same way. Since there can be no doubt that the agent experiences stimuli in the early stages of perception (Kosslyn 1994), so must the possibilities involved in perception be, in some sense, experienced and so enacted (rather than represented) by the brain. This is consistent with reports in perceptual priming (Friston 2010). Perception turns to be quite literally a process of controlled hallucination—as Ramesh Jain has aptly described it (Grush 1995, Ch. 3).

VI. The Bayesian Sense of "Object"

But whenever a path is taken others must be left behind. The commonsensical view that "object" is a blank term applying to everything there is must be abandoned. Or, better, an ambiguity must be postulated to allow for another sense of "object" which applies to the objects of experience which make up possibilities—what I will sometimes call *virtual objects*. This sense may well be necessary anyway to agree on the exact range of application of the first sense (cf. Lewis 1983, Ch. 2, PS. e). Either way, in this second sense, objects are epistemic artifacts: they are the basic ingredients of experience. I have been assuming that experiences are token-identical with brain states and that their content arises, therefore, from high-level, functional characterizations of those brain states. Hence, virtual objects (and properties) are artifacts for the characterization of brain states. But not any characterization: experiences are brain states characterized in terms of

the patterns of dispositions to interact with the world which an agent being in those states realizes. In short, virtual objects and properties simply encode these personal (as opposed to sub-personal) patterns.

Since possibilities are the gist of cognition, this sense of "object" explains the immanent objectivity of the mental observed by Brentano. Yet, since virtual objects are not given as metaphysical ingredients of the world, existence is not an attribute they essentially have. Or, to put it in a way familiar to Bayesian epistemologists, possibilities are logically prior to reality in the architecture of the mind, the latter emerging from the assignment of credences to the former. As a consequence, Brentano's ontological inference, namely, from the mental's immanent objectivity to the intentional inexistence of the objects of thought, is unwarranted.

The interesting question regarding this sense of "object" is not how some objects can fail to exist, but rather: How can some objects exist? After all, some objects appear to be concrete and material, not just epistemic artifacts facilitating the characterization of brain states. The answer is that the characterization of brain states in terms of virtual objects exploits the interface between the brain and the world, namely, virtual objects provide the rubric for organizing the brain's sensorimotor dispositions (Noë 2004).[15] Of course, when we imagine an object, the object does not exist but we recognize the brain state we are in by our basic, off-line sensorimotor dispositions somehow embodied in the object. Yet other times the external part of the characterization of brain states superposes the actual world. In such cases the objects involved in the characterization exist.

This gives some insight into the question of why when we have access to our mental states it is under their characterization in terms of objects and properties. Most likely a fully satisfactory answer should address issues regarding self-awareness and consciousness that are

[15] Cf. Carnap 1956b.

beyond the scope of this dissertation. Still, the brain evolved as an organ primarily for cognizing its external milieu, as witness exteroception, which occupies the vast majority of the brain's sensory capacity. But if what I have said is correct, cognizing the world is preparing to react to it and so cognition also tells about the state of the brain. Since the world is what we have cognitive access to, the best way to know about our brain states is to direct one's attention to the world. This is a different route to the so-called transparency of self-knowledge: to know what I am seeing, I must simply look (Evans 1982, Moran 2001, Byrne 2010). Or, in other words, when I see such and such (or seem to), and am self-aware, I can (normally) know that I am in the state of seeing such and such.

I hardly need to say that I consider this sense the most important, fundamental sense of "object" when it comes to the study of the mind. In fact, I daresay the coherence of the metaphysical sense of "object" inherent in the *manifest image* is questionable. Ultimately, if one is serious about understanding the relation between the brain and the world, one must get passed the primitiveness of objects in our conceptual structure. The reason is simple enough: that conceptual structure itself is shaped by the very relation between the brain and the world. Yet taking that step remains anathema for hardline realists. I shall not hope to persuade them.

The picture of objects as metaphysical realities, given to us for cognition, is agent-centric and, consequently, misrepresents the brain's relation to the world. Absent cognizers, a rock is no more a rock than two half-rocks glued together by molecular bonds and a car is no more a car than a collection of mechanical pieces organized in peculiar relations. It is not that our perception of rocks and cars are sheer illusions. Rather, the point is that, given some underlying physical reality inaccessible to our perceptual system, there is nothing, in the absence of cognizers, to ground the robust unity of any mereological parts of this reality required by the metaphysical

notion of object.[16]

True, to accept that there is a world is to accept that there is something—even if it is no *things* (e.g. Ladyman & Ross 2007). But whatever the world is at bottom (or at a sufficiently deep microphysical level, should there be no bottom) there is nothing to favor any partition of it over others. Worse, contemporary physics teaches that our preferred way of partitioning the world collapses at a sufficiently deep microphysical level. And because no partition of the world has any special status, there are no types that are inherent to the world. This is not to say that the world is composed only of tokens. Rather, the very distinction between *type* and *token* is meaningless in the absence of cognizers.

But imagine that a part of the world, call it P, is such that it reacts selectively to parts of the surrounding world so as to minimize its internal entropy (or maximize its free energy; Friston 2010). To describe P thus involves partitioning and typing it into subparts so that its reactions can be classified as of the same or different types. This internal partition and typing induces a partition and typing of the surrounding world according to the part's responses to it: two parts of the surrounding world are of the same type if P reacts in the same way to them. Assuming that the second law of thermodynamics reflects something fundamental about the world, P is a singularity and, hence, the partition it induces is salient—yet it still lacks any special metaphysical status or unity. To use a familiar metaphor, an absolute cognizer such as God would have no reason to view the world as partitioned that way as opposed to any other. But there are no absolute cognizers and all the cognizers that exist are just such parts of the world as P. Relative to P, P is obviously special.

To say that cognizers are just such parts of the world as P is to say that the brain exists as

[16] Obviously I share my revisionism of the manifest image with Ladyman and Ross (2007). But it should be equally obvious that I disagree with them on some important points. Particularly important in the context of this chapter, I reject the view of modality as inherent in the world, as required by their *ontic structural realism*.

a component of an entropy-minimizing process (Friston 2010). The brain controls a certain class of P's reactions to the surrounding world and so partitioning and typing the world is inherent to its activity. In simpler words, to regard a part of the world as an object is to react differentially to it (cf. Noë 2004). We can call these parts of the world objects, but doing so does not endow them with metaphysical unity independently of us. Perhaps it would be more accurate to say that often some members of this partition of the world realize a model encoding the brain's reactions. This model is the output of a process of browsing through various models in a recursive action-perception loop with the environment. I say that it is the basic ingredients of these models that we usually call *objects*.

The non-arbitrariness of the partition of the world induced by a cognizer relative to its vantage point is important to distinguish the proposal from various forms of ontological relativism (e.g. Putnam 1983, 1987). The parts of the world that fall within this partition vary greatly. But, generally, they tend to be portions with space-time stability. Obviously not every mereological part of the world is a member of the partition. In particular, we can surely rule out the mereological sum of the Eiffel Tower and Bill Clinton's nose as a member (Hirsch 2002). Moreover, while there is no fact of the matter whether a rabbit or its undetached parts have metaphysical priority, there is an epistemic fact of the matter whether the mental models of an agent in the rabbit's surrounding contain one object corresponding to the rabbit or several objects corresponding to the rabbit's undetached parts and so whether the agent's use of "gavagai" refers to the rabbit or to the aggregation of its undetached parts (Quine 1969).

This account of objects is obviously at odds with naive realism about perception (Armstrong 1961, Huemer 2001). But nor is it a plea for sense-data (Russell 1912, Moore 1953). For one thing, virtual objects are not what accounts for our basic perceptions (or misperceptions)

but rather constitute the content of the manifold experiences associated with a single act of perception. Surely we experience edges and shades when we perceive but we also experience *other* virtual objects cropping out of the first ones (Marr 1982). For another, sometimes we may experience objects without perceiving them. I can recreate the experience of being chased by a bear without thereby experiencing perceiving the bear. In doing so, I may experience my desperate running, the terror, the urgency to find a tree around me, etc. Yet, though unperceived, the bear is part of the content of my experience all the same.

In fact, the perennial dispute between naive realism and sense-data exemplifies the conundrums that result from taking the manifest image metaphysically seriously. If objects are given to us to perceive, they must lie somewhere. But the world is too far—on pain of estranging illusions or hallucinations. And the mind/brain is too close—on pain of solipsism. And there seems to be nowhere between to place them. Rather than facing such an uncomfortable choice, my proposal abandons the manifest image as a theoretically reliable conception of reality. Objects are not given to us to cognize: they are byproducts of cognizing the world. We perceive neither sense-data nor fundamental, self-standing realities. We perceive the world—organized into objects by us. Of course, if what I have said above is correct, the ability to organize the world into objects is both automatic and, in large measure, innate. This goes a long way to explaining the compulsory appearance of the manifest image. In fact, it is also partially learned. Born-blind people who gain vision later in life have to painstakingly learn to organize their visual experience into objects. And, as it turns out, early enough so must we all, for vision materializes after birth. We couldn't possibly remember, but this suggests that when we learn to see we are learning to organize the world into objects.

Chapter 6

INTENTIONAL RELATIVITY

I start from the assumption that all agents view the world and its possibilities as composed of, on the one hand, objects and, on the other, properties and relations among them. The objects that compose an agent's view of the world and its possibilities need not be the same as those of another for their life experiences may differ enormously. Yet they may overlap. My goal in this section is to develop an account of how an agent can directly use the objects they themselves recognize for modeling another agent's decision making.

I do not assume that the objects that compose an agent's view of the world and its possibilities exist. Nor that the agent takes them to exist. When the topic is mutual psychological understanding, given that we think and talk about all sorts of objects that don't exist, ontological questions are tangential.

I will assume that psychologically interpreting another is just understanding her decisions. My proposal is not about psychological explanation in general, however, as not all instances of psychological explanation count as psychological interpretation. "David opened the door intentionally" is arguable a psychological explanation of why the door was opened, namely, because David had the intention to open it. But it is too rough an explanation to advance our understanding of David's psychology. Hence, I do not consider it an example of psychological interpretation. Intuitively, psychological interpretation aims at unveiling the reasons for somebody's action.[1]

I have been inspired to develop this idea by the growing conviction among philosophers and cognitive scientists that the computational paradigm of the mind is broken beyond repair and that the field is open to alternatives. My proposal will not resonate with those who see little

[1] In Davidson's terminology, it aims at *rationalizing* an action (1963).

wrong with the computational paradigm (Pylyshyn 1984, Fodor 1987). Nor will it resonate with those who believe in any intermediate realm of the mental such as Fregean senses (e.g. Forbes 1990). Nor will it resonate with those who believe that thought is a linguistic phenomenon (Sellars 1956). In dispensing with intermediaries, my proposal defies all these assumptions.

I will also assume that agents' decision making is captured by the standard formulation of decision theory (Jeffrey 1983). I trust that recent developments of this theory, as those derived from the Newcomb's problem (e.g. Nozick 1969, Lewis 1981b, Joyce 1999), can be incorporated without major alterations. In using decision theory, however, I do not assume that agents are perfectly rational (Tversky & Kahneman 1974). This is a proposal about how we go about understanding other agents when we do; no claim is made that it is always possible to do it. Still, the instances of irrationality are doubtless dwarfed by the instances of rationality even in the most irrational of agents—if they are to be understandable (Davidson 1973b). Moreover, even when agents are less than perfectly rational their failures can usually be incorporated into a decision-theoretic framework (see above discussion of akrasia).

By the lights of decision theory, the only things we need to know about an agent in order to model her decision making on a particular occasion are her utility assignments to the relevant options and her current goals or intentions. In other words, to understand an agent one must have an idea of the agent's basic system of values and credences. These are represented in decision theory by the utility and credence functions.

We should not think of the utility and credence functions as anything more than convenient abstractions, however. For one thing, agents do not literally apply a function to their thoughts when deliberating in anything like an automatic procedure. Usually agents have to labor their way to evaluations through inferential processes. In fact, it seems pretty clear that there is

no unitary neural system in charge of evaluations of possibilities. To be sure, emotions play an important role as evidenced by the involvement of the amygdale and other emotion-related areas of brain in decision making (Damasio 1994). But, at least in humans, it seems at best a partial role as it can be snatched by activity in the prefrontal cortex believed to be associated with planning and the calculation of consequences (Greene et all2001). Moreover, the idea of assigning numerical values to options is, too, a mere theoretical artifact. Usually, agents simply rank options pairwise among those relevant to their current deliberation. For this reason, a preference relation among options is probably more apt to capture what goes on in people's heads than a utility function. However, when an agent's preference relation is acyclic and total, a utility function can always be defined which perfectly emulates it (Sen 1997). People's preferences vary nearly all the time, but considering time slices of people, the assumptions of totality and acyclicity seem to me as good an approximation as the attribution of rationality. Either way, what I have to say will not crucially turn on the characterization of preferences and credences as numerical functions.

I. The Formalization

Like the Simplest Quantified Modal Logic (Zalta & Linsky 1994, Williamson 1998, Menzel 1991), let us begin with a set of objects such that they do not have any properties essentially except their identity—call it the agent's framework of objects or O for short. We can think of the elements of O as bare possibilia or, as it were, hangers of properties essentially different from one another (Priest 2005). But instead of introducing possible worlds as primitives and defining relations in terms of them, as is typically done, let us take a set of atomic relations (including unary relations or properties) also as primitive—call it R. N-nary members of R combine with n-tuples of members of O to form what I shall call, in Russellian style, atomic propositions. Any

combination of an n-ary relation in R with an n-tuple in O, for any n, is an atomic proposition, and I will call the totality of them the agent's epistemic space. Next, I will adopt the standard definition of truth for Russellian propositions (Soames 1999): '$R_n(o_1, \ldots, o_n)$' is true iff $R_n(o_1, \ldots, o_n)$. (This, of course, requires that $o_1, \ldots, o_n$ exist.)

Let us also define the logical connectors ~, ∧ and ∨ (viz. negation, conjunction and disjunction, respectively) defined over the agent's epistemic space in the standard truth-functional way. The negation, conjunction and disjunction of propositions are also propositions.

Also, let us provide the agent with an unlimited supply of auxiliary objects, *x*, *y*, *z*, etc.[2]It is appropriate to call these elements objects because they can combine with elements of R in the same way that members of O do. However, these objects are not themselves members of O because their identity is indeterminate. This does not mean that they are not necessarily identical to themselves and nothing else. Rather, it means that which object they are is either unknown or irrelevant. In few words, like variables in first-order logic, auxiliary objects supplant objects, both in O and others. But, like expressions containing free variables, the combination of auxiliary objects with relations does not yield propositions unless within the scope of the quantifier ∀, defined thus:

$\forall x\alpha(x)$is true iff$\alpha(x)$ is true when x supplants any object (whether or not in O).[3]

Finally, let us introduce the possibility operator ◊ in the following way:

$\Diamond\alpha$ is true iff α is logically consistent,

[2] Auxiliary objects are inspired in the idea that the elements of mental models do not wear their identity on their sleeves in that we can conjure up all sorts of models with indeterminate elements (See chapter VI section).

[3] Beside the general and particular quantifier of first-order logic, which can be defined in terms of one another, it is easy to define ordinary language ones such as "the such and such" associated with definite descriptions (Neale 1990) or others (see Barwise & Cooper 1981).

where α is any Russellian proposition formed with the above elements. Of course, $\Diamond\alpha$ is also a proposition.

Finally, take the closure of the agent's basic epistemic space under $\sim$, $\bigwedge$, $\forall$ and $\Diamond$. The result is a σ-algebra, call it Ω, whereupon we can define the agent's utility and credence functions in such a way that they satisfy the following constraints:

$$C\left(\bigvee_{i=0}^{\infty} p_i\right) = \sum_{i=0}^{\infty} C(p_i)$$

for any $p_1, p_2, \ldots$ mutually exclusive and $C: \Omega \rightarrow [0,1]$ is non-null; and

$$U(A) = \sum_{i=0}^{n} C(S_i|A)U(S_iA)$$

for $\{S_i, \ldots, S_n\}$ partition and $U: \Omega \rightarrow \Re$.[4] I will call the tuple $< U(\cdot), C(\cdot) >$ the agent's worldview.

II. Some Benefits

This formalization of decision theory differs from Lewis' formalization (1981b) in terms of possible worlds in a few important respects. Firstly, taking the ingredients of deliberation to be composites of objects and properties has the advantage of accounting for what we may call *mental particularization*, namely, the inference validly allowed by the use of the word "about" from an attribution of propositional attitude to the attribution of an objectual attitude. For example, in common parlance we transit with ease from statements of the form "S has a V that p" to statements of the form "S has a V about x," where "V" is any propositional-attitude verb and

[4] Conditional credence is defined as standardly done.

"x" is a referential expression appearing in "p."[5] Because Lewis' propositions are equally associated with scores of objects, namely, those populating the possible worlds that belong to the propositions, mental particularization is mysterious on his account. If I believe that Obama is the US president, I presumably bear the same relation to a vast number of whole possible worlds, each one containing scores of objects. Hence, I do not seem to bear any special relation to Obama or the US presidency on account of that belief and so there is no obvious reason why my belief should be about Obama as opposed to, say, Obama's father, planet Earth or space-time. In contrast, on the Russellian-style view proposed here mental particularization is natural, as it allows only the items participating in the proposition involved to bear a relation to the agent in virtue of her having an attitude toward it.

Secondly, since Lewis' possible worlds are ontologically independent categories, he has no choice but to require that the utility and credence functions be defined over the entire domain of possibilities. This is counterintuitive, however. Did Plato assign any credence to the proposition that Obama plays basketball every morning? This seems highly questionable, for he obviously lacked the concepts of Obama and basketball. Of course, there is a fairly standard sense in which an agent is ignorant of a proposition when she has a degree of confidence of 0.5 that it is true. However, a desideratum of an account of decision making is that it distinguishes between, on the one hand, an agent being torn between a proposition and its negation and, on the other, an agent assigning no credence at all to a proposition and its negation because she lacks the conceptual resources to even entertain it. Likewise, a desideratum of an account of decision making is that it distinguishes between an agent being indifferent between a proposition and its negation, and her assigning no utility to it whatever because the proposition is estranged from her conceptual repertoire. These desiderata are captured in the requirement that the credence and

[5] Notably: even when the referential expression is empty and so the object involved does not exist.

value functions go undefined over propositions that cannot be formed with the agent's conceptual resources. This requirement is elegantly satisfied by my proposal, for an agent's utility and credence functions are defined over the propositions composed of only objects whose identities are contained in the agent's framework of objects, O.

There is also some empirical evidence that the elements of one's worldview feature importantly in one's psychological understanding of others. By the age of 4, children finish the development of a full theory of mind, that is, they finally master folk-psychological explanation of others. The last hurdle in this development is their learning to attribute false beliefs, as testified by the so-called false-belief task (Wimmer & Perner 1983). However, studies suggest that children have a pretty good understanding of agency from much earlier (Gergely & Csibra 2003) but that they are unable to inhibit their own worldview in psychologically interpreting others. This is why they can't incorporate false beliefs into their psychological understanding of others (Leslie, Friedman & German 2004). When a child less than 4 years old sees the girl go away and the lab assistant change the girl's doll from the hole where she left it to the other (as in the classic false-belief task) the child's psychological explanation of the girl must fundamentally rely on these objects if her failure consists in her inability to inhibit her own knowledge of the doll's location. What they learn at the age of 4 is, therefore, to assign properties to the objects of their own worldview which she does not believe them to have. In short, what they learn is to construct different worldviews with the elements of their own.

III. Decisions

We need not have an exhaustive knowledge of an agent's worldview to psychologically understand her. In fact, it seems nearly impossible to have such knowledge of any agent (including ourselves, if Freud was right). It is more plausible to take the notion of understanding

an agent as relative to contextual standards of thoroughness. Yet decisions seem to be the units over which these standards trade. We understand an agent, to whatever degree, when we understand the decisions she makes. And, since I am assuming that agents are utility maximizers by constitution, we understand the decisions she makes when we understand the decision problems she faces.

Following Joyce 1999 (Ch. 2), let us think of a decision problem as a structure:

$$D =< I, A, S, R, C_s(\cdot\,|A), U_O(\cdot) >$$

where *A*, *S* and *R* are subsets of the agent's σ-algebra Ω. $U_R(\cdot)$ stands for the agent's utility function $U(\cdot)$ over *R* or, what comes down to the same, the assignments of utility to outcomes by $U(\cdot)$. Also, let $C_S(\cdot \mid A)$ be the set of credence functions over *S* resulting from $C(\cdot)$ conditionalized on each member of *A* or, what comes down to the same, the assignments of credences to conditions in S given a proposition in *A*.

A stands for the set of actions the agent can choose to try to perform and *I* stands for the immediate intention or goal the agent has in performing an action in A (see chapter 2). I interpret a choice of action as involving only commitment on the part of the agent to investing her resources to make the proposition true. I want to allow for cases where the agent does not even believe it likely that she will succeed: e.g. a normal person trying to score from the middle of a basketball court. S is a partition, that is to say, a set of propositions such that exactly one must be true.

The propositions in *S* represent conditions that the agent does not purport to control—at least for the purpose of this particular decision.

Finally, *R* stands for outcomes (or results) and includes propositions such that the agent

assigns a high credence to their truth given the truth of combinations of propositions in *A* and *S*. Intuitively, the agent views the propositions in *R* as consequences of the conjunctions of her actions and the conditions given in *S*. There is, therefore, some redundancy in defining the problem. In fact, some go as far as to identify the members of *R* with conjunctions of members of *A* and *S* (Jeffrey 1983). This redundancy is welcome, however, since so defined decision problems incorporate the idea, central to our notion of agency, that the unfolding of events is the resultant of the independent contributions of the world and the agent (e.g. Belnap, Perloff & Xu 2001).

In one sense this individuation of decision problems is restrictive, for it does not allow two agents to face the same problem unless their utility functions and their credence functions, conditionalized on *A*, coincide over *S*—even if the situation they are in is externally identical. This is not incompatible with the recognition of various similarities among decision problems, however. Decision problems are portrayed as manifold and, accordingly, can share numerous properties, hence belong to numerous interesting types.

Yet in another sense this individuation is permissive, for it allows two agents to face the same decision problem even if their utility and credence functions differ everywhere else than over *S* and *R*. The utility and credence functions are idiosyncratic and so, if the individuation demanded that two agents have the exact same utility and credence functions in order to face the same decision problem, then no two agents ever would. This seems to me an undesirable consequence. In the end, the individuation of decision problems here proposed is custom-made for the intuition that a necessary condition for two decision problems to be the same is that they have the same solution, hence that they yield the same decision.

IV. Mental Coordination

Psychological explanation has tacitly been assumed to be a garden-variety explanation in that the explanandum is a phenomenon given to the cognizer for her to understand.[6] The agent's mind has been conceived as an independent reality that we unveil when we come to understand the agent psychologically. This conception of psychological explanation forces a conception of the mind according to which the physical reality suffices to fully determine it. Consequently, to accommodate mental phenomena which do not harmonize with reality, such as beliefs about fictional objects, Frege-style puzzles, perceptual illusions, etc. philosophers are forced to postulate intermediaries that provide an extra degree of freedom. Thus, abstract objects, senses, modes of presentation, languages of thought, sense-data, etc. have been born. With an extra variable at their disposal, mismatches between the mind and reality can be chalked up to it. Yet more than a century of unsuccessful attempts to fit this extra degree of freedom in the remaining picture should give us pause (see Chapter 7).

My proposal opens the door to a radically different way of thinking about psychological explanation. Rather than a one-mind phenomenon given to the cognizer, I believe that psychological explanation may irreducibly involve the coordination of two minds. This is at the heart of the idea that an agent can use the objects, properties and relations that shape her own worldview to model the decisions made by another. At a minimum, it requires that interpreter and interpretee share some of the objects (and properties) that frame their worldviews. A precondition for my understanding your actions is that I associate a set of objects with you. But this requires me to recognize you as well as those objects in the first place. In other words, to view certain objects as the locus of reference and thought for you, so must they be for me and so must *you* be for me. I ought, therefore, to view you as a special object in my objectual framework in that you are attuned to other objects therein. When this happens I view you as

[6]With the exception of the imprecise views of Dennett (1987).

sharing a part of my worldview and so can you view me as sharing your worldview. This shared mental space acts as a frame of reference for coordinating our actions.

This shared mental space is not bound to reality: our mental coordination almost always surpasses the perceptually accessible and often the actual. It is not, therefore, required that the objects that participate in the psychological interpretation of you exist; nor that I, the interpreter, take them to exist. More generally, it does not require that your worldview (or mine) harmonizes with the independent reality as long as I can synchronize it with mine.

This also explains the non-epistemic appearance of possibilities, even if they turn out to be fundamentally first-personal. By bridging worldviews, psychological interpretation is the glue that allows us to construct an intersubjective modal space: a domain of objects, actual and merely possible, which we see ourselves and our countrymen as inhabiting (see Chapter 5).

V. Folk-psychological Notions

One's knowledge of the objects that framed another's decision may be partial, thereby preventing a full grasp of the decision problem she faced, yet still allow for some understanding of the decision. I may rightfully say that I understand Caesar's decision to cross the Rubicon with his army forcing thereby a civil war on the basis of, first, his assigning a high credence to the proposition that, had he not done so, he would have been accused of treason upon his return to Rome and, second, his valuing this outcome more negatively than a civil war. But it would be unjust to Caesar to suppose that this description of his decision reflects the actual cognitive process he underwent. Surely he contemplated more options than just crossing the river with his army or returning to Rome alone, and probably these options broke down into more specific ways of doing either. Perhaps he contemplated exile or a coup d'état. The point is that oftentimes understanding a decision consists in having a grasp of those elements of the decision problem

faced by the agent that are explanatorily important.

In fact, this flexibility in our understanding of decisions even while ignoring the exact decision problems that yielded them might be the reason why we do so well with a system of discrete notions such as beliefs and desires, despite the putative need of degrees to scrutinize most actions. When we say that Caesar believed that he would be beheaded if he returned to Rome, we may simply be saying that he assigned high enough credence to this proposition to dominate the explanation of his actions. If he had had the same degree of conviction on the truth of this proposition yet had returned to Rome without his army out of fear for the life of his wife, we would explain his action as the result of his conjugal love instead. Perhaps in that case it would ring appropriate to say that Caesar feared he would be beheaded in Rome rather than believed it, despite the fact that, *ex hypothesi*, he assigned the same credence to it as before. The difference lies in how this proposition affects the comparison between the expected utilities of Caesar's options and so in the explanatory weight it carries in both cases.

This somewhat deflationist account of our basic folk-psychological notions has the advantage of construing them as essentially non-occurrent states. There can be little doubt that our common parlance favors this interpretation: *Why does John continue to quietly read his book instead of running out of the house? Because he believes that the house is sturdy*. Yet since the credence and/or utility assigned to a proposition must generally be high for it to dominate the explanation of an action, we should all the same expect attributions of beliefs and/or desires to be reliable clues to an agent's permanent psychological make-up.

While philosophers have tried to rest much of the weight of an account of practical reason on the concept of reason itself (e.g. Parfit 1984, Korsgaard 1990), the colloquial use of the

term seems amenable to the same deflationary treatment.[7] Thinking of decisions in light of decision theory offers a natural explanation for why the reasons for somebody's actions may be on one occasion a belief and on another a desire: we use the word "reason" to introduce a commentary on a decision problem. More precisely, when A tells B the reason why C did what she did, A is providing B with the most important, salient aspect of C's decision within A and B's conversational context. When the context is purely explanatory, C's reason for acting as she did is the weightiest contributor to the utility differential between the options or the aspect most informative of the agent's situation vis-à-vis A and B. But when the context involves important evaluative elements (e.g. moral judgments) the reason why C did what she did may well be that aspect of her decision that most aptly evaluates her (e.g. moral) character. Either way, if this is correct, reasons are dependent on explanatory context and so arguably lack the kind of metaphysical robustness required to ground an account of practical reason.

How do we go about finding out what objects another recognizes? This is the epistemological side of interpretation which I will now address.

VI. Radical Interpretation

Differences in formulations notwithstanding, we may think of the challenge of psychologically understanding another as the problem of radical interpretation (Davidson 1973b). In full generality the problem is this: given all the information about the physical history of an agent (including causal transactions with her environment and functional facts), what mental states am I to attribute to her? (Lewis 1974)[8] Or, in decision-theoretic terms: given all the physical history

[7] Here I am referring to internal reasons (Williams 1979).

[8] Lewis took psychological explanation as the cognition of phenomenon given independently of the cognizer (see above). Though not much depends on this in the remainder, here I use the pronoun "I" to mark a difference with Lewis' formulation. I don't just mean "I" as a stand-in for any person, regardless of her own worldview, but rather as a more intimate representation of her psychological individuality.

of an agent, what utility and credence functions am I to attribute to her? A function is equally composed of domain as of image (and, of course, the pairings between them) and so part of attributing utility and credence functions is to attribute a domain of propositions over which they range. As we saw, this is a product of both a framework of objects as well as a set of properties and relations. Though these are inseparable in practice, here I shall focus on the attribution of a framework of objects.

The principal advantage of formulating the problem of radical interpretation in terms of objects is that it dispenses with the middle man. Language has taken center stage in previous formulations of the problem of radical interpretations (Davidson 1973b). And even when the problem is explicitly formulated as being about mental states rather than language, it is granted that solutions must be given in our language or the agent's (Lewis 1974). But language is an unnecessary distraction. Moreover, its disruptive presence excludes nonlinguistic creatures from those who can take part in interpretation at both ends. In contrast, I start from the assumption that all agents view the world and its possibilities as occurring within a framework of objects and that any agent can use her own framework directly (provided she recognizes other agents as such) to model the way others view the world and its possibilities. Among these objects may be sentences and among the relations may be intentional ones and even conventions (Lewis 1969), hence interpreting language is but a sub-problem of the problem of radical interpretation, namely, the problem of figuring out an agent's intentions in uttering a given sentence (Grice 1975).

As suggested in chapter 5, perhaps these frameworks of objects are ways of organizing and characterizing experience; or perhaps they are something else, though they show up in experience. But whatever they are, they are, in some sense, real and frame worldviews which involve painstaking detail. Explaining interpretation as the synchronization of worldviews makes

the first- and third-persons immediately commensurable exposing the magnitude of the theoretical problem of radical interpretation as formulated. A solution must be such that it fully superposes the agent's own worldview. This provides much relief against the woes of indeterminacy (Lewis 1974): not too many utility and credence functions will superpose an agent's worldview, hence there cannot be many solutions to the problem of radical interpretation.

But the point of the theoretical problem of radical interpretation is simply to understand the dependency of attributions of mental states on the physical facts about the agent and her environment. In practice, of course, it is impossible to solve the theoretical problem of radical interpretation for any agent. In fact, abandoning the assumption of physical omniscience reveals the true importance of language for the practical problem of interpretation. For, under normal circumstances, an agent's worldview (e.g. her perceptions, experiences, etc.) are completely inaccessible from the third-person. We accordingly face great epistemic indeterminacy, namely, the evidence accessible from the third-person may not settle which of various hypotheses best matches an agent's worldview. But language can partially alleviate the problem. When interpreter and agent possess a language in which to communicate, we should expect the agent's speech to carry much information about her worldview facilitating interpretation (Davidson 1973b).

But there is an irreducible, non-epistemic type of indeterminacy in the problem of radical interpretation. Even if a framework of objects with their associated properties and relations matches an agent's worldview seamlessly, there is still the question about the identity of the objects. Recall objects do not have any property essentially except their identity and so, given the right cardinality, any set of objects should be as good as any other to serve as the framework wherein to model an agent's worldview and particular deliberations. In other words, not always

does an agent's worldview settle the identity of the objects that act as a framework for it. Ordinary cases where this comes up typically involve misidentification but the most vivid illustration comes from a rather extraordinary case: the so-called slow-switching cases derived from Putman's famous Twin Earth thought experiment (Putnam 1975).[9]

Oscar, the protagonist of Putnam's thought experiment, is an inhabitant of Earth and his doppelganger, Twin Oscar, is an inhabitant of Twin Earth which is an exact replica of Earth in every observable respect. Imagine, further, that by the work of some science fiction-like technology, the Oscars are instantaneously transported to each other's places in Earth and Twin Earth, respectively. Intuitions, it is generally agreed, suggest that initially Oscar's thoughts will be about the objects of Earth but after some time in his new environment they will become about the objects of Twin Earth (Boghossian 1989, Ludlow 1995a).[10] If we take these intuitions seriously, we should ask: at what exact point in time does Oscar's belief that water is liquid becomes the belief that *twin water* is liquid? Yet there seems to be nothing in the metaphysics of the situation to answer this question. Consequently, either the change is vague, somehow gradual or conventional. Either option entails that there is irreducible indeterminacy in radical interpretation.

The intuitions that sustain this argument arguably concur with the intuitions appealed to by Descartes' skeptical conclusions. Therefore, a credible account of psychological interpretation had better explain this irreducible form of indeterminacy. The question is: how—without making psychological interpretation arbitrary? Here is a recipe: find two principles governing the attributions of objectual frameworks such that when these kinds of conflict arise they are settled by the principles' relative weights. Then explain slow-switching cases and the like as cases

[9] This is the problem Searle calls the problem of particularity (1983, Ch. 2).

[10] In fact, we can switch the Oscars back to their respective planets and the same intuitions mandate that the same effect should ensue, namely, the contents of their beliefs should switch, after enough time, from their previous environment to their original one.

where the passage of time gradually shifts the relative weights of these principles. This is what I plan to do next.

VII. Intentional Inertia

The first principle derives from a constraint of global coherence, which is a somewhat stronger version of the previously mentioned constraint that attributions of utility and credence functions superpose the agent's worldview. Objectual interpretation is inspired by the intuition that psychological interpretation consists in using the elements of one's worldview to model another's worldview and thereby her deliberative processes. If so, our assignments of objects to an agent must correspond one-to-one with her own framework of objects, for it is essential to how agents view the objects that frame their worldview that these objects quite directly connect them to the world and its possibilities. Or, conversely, one cannot suppose that the objects of one's worldview present to one more than one object each, for this would be tantamount to supposing that one's worldview is flawed. This entails what I shall call the principle of intentional inertia: *once an object of our own has been assigned a role in the agent's worldview neither assign it other roles nor assign other objects to that role*.

For example, if little David is first told about God by his parents and later, after being told about Santa Claus, he takes Santa to be God, we would not say that David's concept of God is about both God and Santa. Rather, it is natural to say that it is about God but that he thinks Santa is God. Translated to my account, in David's framework of objects we must include only one object in place of God and Santa for there is only one role associated with both. Which one? The principle of intentional inertia mandates that since God had already been assigned to David's worldview, no other object is assigned to it and, accordingly, it is God who stays and Santa who

is left out.[11]

Cases like these in the literature are said to raise the disjunction problem (Evans 1982, Fodor 1990). The problem is typically formulated in terms of a frog which snaps at flies in her natural habitat. But if we put the frog in an environment that replicates her natural habitat except for the fact that, in addition to there being flies, there are identical-looking placebos the frog also snaps at, why not say that the frog's concept of fly is about both flies and placebos? According to my proposal, the answer lies in the one-to-one correspondence constraint on interpretation. If we take it as the baseline case that the frog snaps at flies, then when she confuses the placebo with a fly, this no more makes her concept of fly about the placebo than David's confusing God with Santa, makes her concept of God about Santa. As will emerge, we might, under special circumstances, decide that her concept is about the placebo but then it would no longer be the concept of fly.

Much more prominent are the inverse cases where an agent thinks one object to be two (Frege 1892, Russell 1905b, Quine 1956, Perry 1980, Schiffer 1987a, Salmon 1986, Richard 1990, Recanati 1993, etc.). Consider Quine's Ralph who after meeting Ortcutt, an important man in the community, thinks him a different person (in fact, a spy) when he see him at the beach wearing a brown hat. Intuitions strongly suggest that we individuate Ralph's mental states more finely than the part of the world they are about or, in other words, that he has two concepts corresponding with one object in the world. This leads to all sorts of complications involving descriptions of his mental states. Frege first proposed the historical case that Babylonians, though otherwise sophisticated astronomers, thought that Venus was actually two planets, Phosphorus and Hesperus, based on independent observations. His solution to the

[11] Still this doesn't mean that God cannot have properties possessed by Santa and not by God, since objects do not have any properties essentially.

aforementioned complications was to claim that when we talk about mental states we are not talking about objects (that is, ordinary objects) but rather about a separate mental realm of senses which is more fine-grained than the world. Almost invariably since Frege, philosophers have followed his strategy of positing a more fine-grained, intermediate realm between the world and the agent.[12] The debate has largely been reduced to which intermediate realm is more metaphysically respectable (see Chapter 7).

Yet by and large these philosophers have failed to notice that this phenomenon and the one that generates the disjunction problem are two sides of the same coin. Worse, intermediate realms are useless in the latter cases because the problem with them is that the agent's mental life is more coarse-grained than the world. My proposal appeals to the same principle to rule both. In Frege-style cases, the one-to-one constraint on interpretation demands that we include in the agent's framework of object some extra ones, for her worldview has objects absent in ours (and, therefore, the world, as we see it). What objects are we to include is irrelevant as long as they are not ones already assigned to her framework. For simplicity, we may use auxiliary objects whose identity is indeterminate besides being distinct from other objects of her framework. In Ralph's case it is natural perhaps to introduce only one extra object playing the role of the man Ralph saw at the beach and let Ortcutt in too. But other times it may be more parsimonious to introduce two auxiliary objects to take the place of one of our own. In the case of the Babylonians, for example, since both Phosphorus and Hesperus seem to have equal right to occupy Venus' role, it feels more natural to introduce two auxiliary objects for them and simply take Venus out. However, this may be a product of our ignorance, for had we been present when the Babylonians first spotted Venus as, say, Phosphorus, the principle of intentional inertia would have compelled

[12] I count neo-Russellians in this group because, even though they consider the aforementioned intuitions as pragmatic (as opposed to semantic), they do not deny them. Instead, they, too, account for them based on an intermediate realm: i.e. sentences (Soames 2002) or ways of thinking (Salmon 1986).

us to keep this assignment.[13]

I do not rule out that there are other, more refined interpretative principles that govern these choices or even that there is yet another type of indeterminacy in these cases. But either way, I will not explore this issue, for once we acknowledge that an agent's worldview is flawed any hope of perfectly synchronizing our worldview with theirs is lost. In effect, acknowledging that an agent's worldview is flawed in a specific way is nothing but recognizing a mismatch with our own!

VIII. The Embedding Principle

The second principle governing the attributions of objectual frameworks I shall call the embedding principle. We have no reason to be interested in other agents' mental life per se. Our interest in other agents is the result of the peculiar nature of their interaction with their environment and the tremendous impact that this interaction has in our lives.[14] There can be little doubt that the ability to predict other agents has an enormous survival advantage (evident both in the predator-prey relation and in the benefits of cooperation) and this is one of the main forces that drove the explosive evolution of the brain (Parker 2003). Consequently, the ultimate reason for keeping track of another agent's mental life is to map out and predict her interaction with her environment and thus the impact on our goals and plans. It is natural, therefore, that we, interpreters, seek to view an agent's mental life as accounting for much of her causal interaction with her environment because the more systematically it does, the more reliable our predictions are. The embedding principle captures this natural inclination by mandating that when an agent acts upon her environment we frame the deliberation leading to her action in the objects of that

[13] This intuition seems to be also at work in the causal theory of reference.

[14] I don't want to exclude the possibility that part of our interpretative responses are innate. Even if they are, the point stands that, most likely, they evolved to raise our chances of survival (e.g. Wilson & Wilson 2003). But not much turns on this point.

environment. This, of course, translates into pressure to incorporate those objects into the agent's worldview.

Unlike the principle of intentional inertia which derived from a global constraint on psychological interpretation, the embedding principle is local in that it derives from the situated import of every instance of psychological explanation. The contrast can be illustrated with an example which also suggests that the embedding principle is wired deeply in our brains.[15] In experiments (Gergely & Csibra 2003) children as young as 1 year of age were shown to attribute agency to a geometrical figure as simple as a circle when its pattern of movement minimized its projected distance from a moving target (cf. Pac-man). Children as well as adults view the circle as *chasing* the target and, accordingly, as tracking its position which translates into a natural inclination to view the circle as an agent of sorts and to incorporate the target as an object of its worldview. This is what the embedding principle captures (though the latter inclination is defeasible if, for example, we knew that the circle is conflating the target with another object). But the case shows that the inclination arises even when we have no clue as to the alleged agent's overall psychology or its perspective to the world. For what can a circle's worldview be? The embedding principle captures, therefore, our inclination to characterize an agent's psychology based on segregated situations and in this sense is local.

The embedding principle explains the intuitions that have motivated the doctrine known as externalism about mental content or, for short, meaning externalism (Putnam 1975; Burge 1979, 1986). The physical variety of meaning externalism, introduced by Putnam's Twin Earth thought experiment, is straightforwardly explained by the embedding principle. A major advantage of developing psychological interpretation relative to an interpreter is that it allows us

[15] Other principles apply to particular types of agents depending on, for example, their perceptual system. The principle that mandates to incorporate an object into a predator's framework when she focuses her gaze on it does not apply to all animals on the prey side of the food chain because they lack frontal vision.

to explain Burge's social externalism, too, with the same principle. Recognizing that I am a member of a culture, I may rely on conventions of various kinds (Lewis 1969) to fix the content of my mental states. For one, I may surely rely on expert knowledge to determine the exact kind that occupies the role of arthritis in my mental states—what Putnam aptly calls division of linguistic labor. But, for another, I can rely simply on intergenerational conventions to fix what object I do not believe to exist when I don't believe in God. In fact, these are the same conventions that determine that Santa and God are distinct objects so that little David can be confused as to their identity. Participating in a culture is not just abiding by certain linguistic conventions or endorsing certain values: it is also (and perhaps more fundamentally) chiming in a domain of objects and possibilities made possible by mutual psychological understanding.

The reason why we view agency in the circle is that its behavior shows a pattern of maximization suggestive of an intention or goal (Gergely & Csibra 2003). Had the circle had the intention to catch the target but failed to fulfill it because, for example, it could not track the target, we would see no sign of agency and, hence, the embedding principle would not exert its force upon us. This suggests that the force of the principle admits of degrees according to the agent's amount of success. The more successful an agent is in navigating her environment, the greater the inclination to view the objects that compose it as part of her worldview. In fact, the force of the embedding principle derives from an inference to the best explanation: the more systematically an agent's plans and intentions are fulfilled by her actions, the better the explanation that she was acting upon a true map of the relevant parts of her environment becomes. By contrast, when an agent's plans and intentions are unsuccessful a salient explanation, other things equal, is that she is misrepresenting the world and its possibilities. In some cases there might be some pressure to incorporate the objects of an agent's environment

even when she is less than fully successful, but in the extreme, that is, in Descartes' skeptical scenario (or Putnam's a brain in a vat) there is no pressure at all.[16]

In fact, the same rationale works from the first-person perspective because success and failure are normally conspicuous. When an agent's worldview is importantly flawed, it will usually lead her to failure in her plans and intentions in a systematic way. If, for example, I keep confusing Peter with John I may say to him things that he does not understand, perhaps may refer to him in too familiar a way or have false expectations as to his behavior. Systematic failure will normally arouse my suspicions that something is wrong with my worldview and potentially lead to mental reorganization. In short, when one's worldview is importantly wrong, the world itself will often tell her.

However, slow-switching cases are made so that the world conspires to perpetuate the agent's error. We know that Oscar is (initially) mistaking every object in Twin Earth for an object of Earth. Yet Oscar is as successful in navigating Twin Earth as he was in navigating Earth and so the flaws in his worldview are undetectable from his perspective. But if it makes no recognizable difference to Oscar to be in Earth or in his doppelganger's place in Twin Earth, why do we have the intuitions that his mental states remain about the objects of Earth temporarily to then become about the objects of Twin Earth? A natural conjecture is the increase of the force of the embedding principle versus the constancy of the principle of intentional inertia.

IX. Conclusion

Ex hypothesi, the switch of the Oscars did not alter their worldview and so immediately after the

[16] An interesting test would be to consider the case of a Cartesian victim such that, though the evil genius is implanting all her experiences in her mind, he is doing so in a way that they systematically coincide with the experiences she would have if she were interacting normally with her actual environment. Would we still feel compelled to view her thoughts as about the objects of her environment? The case is also interesting because, were there a consensus on the answer, it would constitute a test for causal theories of content vis-à-vis the embedding principle—that is, provided that the evil genius does not know about the agent's environment through causal means (e.g. Berkeley's God).

switch the principle of intentional inertia dominates psychological interpretation forcing us to not reassign objects to Oscar's worldview that were already assigned. Thus we feel that initially Oscar is conflating every object of Twin Earth with its doppelganger in Earth. However, Oscar navigates his new environment swimmingly and as the successes mount and as the history of his actions and plans becomes more anchored in the objects of Twin Earth, the pressure to incorporate these objects into his worldview increases. Eventually the embedding principle will dominate psychological interpretation and Oscar's mental states will become about the objects of Twin Earth. If we then switch the Oscars back to their planets, the principle of intentional inertia will temporarily fence off the assignments of twin objects to Oscar's worldview until the embedding principle catches up to it. The process repeats.

The same phenomenon does not occur in Putnam's original thought experiment (1975) because the Oscars were never severed from their environment and so the opportunity never arose for the principle of intentional inertia to rise to relevance. The embedding principle dominates the situation without counterweights spurring the intuitions we associate with meaning externalism.

The indeterminacy as to the exact point in time at which the change happens seems to me rather natural. After all, the embedding principle derives from an inference to the best explanation and the question when the embedding explanation is best may depend on a variety of contextual considerations such as Oscar's type and degree of attachment to the persons and things in Earth, the relative importance of the events that took place in Earth in defining him, etc. In fact, there may not even be an interpreter-independent criterion but only the vague assurance that for every interpreter there always is one or another.[17]

[17] I do not rule out that under special circumstances the weight of the embedding principle never catches up to the principle of intentional inertia, e.g. concrete events that took place in Earth occupy a prominent role in Oscar's mental life till his death or he

I do not fear that this would make psychological interpretation capricious and unruly. To be sure, folk psychology is based on principles, the principle of intentional inertia and the embedding principle being but a small sample. But it is a practical theory: its principles evolved to give clear answers in situations where clear answers matter for real agents. It is, therefore, no surprise that they do not deliver sharp answers to philosophers' thought experiments. Rather, the surprise is that they go so far as to tell us this much about them.

dies too soon after the switch.

Chapter 7

OPACITY AND METONYMY

Consider the following version of the principle of compositionality:

(C) Compositionality: The basic proposition meant by a speaker with a declarative sentence is determined by the sentence's structure and the content thereby meant by the speaker with each of the sentence's components.[1]

Though defended from its most prominent objection by Szabó (2010),[2] this principle in full generality seem immediately objectionable given the pervasiveness of underarticulation in everyday linguistic communication, that is, situations in which the basic proposition meant by a speaker with a use of an unambiguous sentence is underdetermined by the sentence's "syntax, the meaning of the morphemes in [it], the prosodic features of [it], the references of any referring expressions in [it] (including those of an indexical nature), and the specification of all anaphoric links" (Neale 2007, p. 252).[3] However, for underarticulation-free linguistic acts (C) does seem quite plausible. There is no prima facie evidence of underarticulation in belief reports and so (C) seems applicable to them.

[1] I will use 'constituent' for a part of a proposition and 'component' for a part of sentence. "the content *meant* by the speaker" with a component of a sentence, X, is intended to capture the commonsensical sense of 'what S was *referring* to with X.' Also, I prefer "meant" to "said" or "expressed" for three reasons. First, I take 'to mean' to be the most fundamental communicative attitude and since I do not plan to say anything original about how communicative attitudes and achievements are related, I take it that I cannot be losing any strength in committing to it. Second, 'to say' and 'to express' have been so abused in the literature that it is no longer possible to use them without suggesting undesired implications. Thirdly, and most importantly, the presence of a speaker is crucial to my argument throughout and to detach the speaker from 'meant' strikes as more unnatural than to detach it from 'said' or 'expressed.' I rely on the strength of this link in allowing myself the liberty to omit the speaker. Finally, I will use 'with X' to associate a linguistic act with the specific expression X, while 'in Y' to associate a sentence or content with the linguistic act Y. I am indebted to Stephen Neale for opening my eyes to the wild inconsistencies in the uses of all these terms in the literature and to the precision that the main argument of this chapter requires. These uses roughly follow his in Neale 2007.

[2] Szabó formulates the principle in terms of "the content of an assertion" instead of "the basic proposition meant…" probably because he conceived it in terms of speech acts. However, this makes the principle indefensible if conversationally implicated contents can thereby be asserted. To exclude contents whose structure departs wildly from the sentence used by the speaker as conversational implicatures permit, I prefer to restrict the principle to "the basic proposition meant by the speaker," which, I think, is intuitive enough but will make precise below. I further restrict the principle to declarative sentences to exclude implicatures not associated with a more basic proposition because they are made with sentences which do not have propositional contents (e.g. questions, commands).

[3] For other advocates of the underarticulation thesis see Sperber & Wilson 1986, Carston 2002a, Bach 1994b, Recanati 2004. For an argument against, or otherwise qualification of, the thesis see Stanley 2000 and Stanley & Szabó 2003.

On the other hand, we have a robust intuition, first observed by Frege, that belief reports such as those made with

1) Hammurabi believes that Phosphorus rises in the morning

and

2) Hammurabi believes that Hesperus rises in the morning

differ in truth value, despite (1) and (2) being composed by only corresponding synonymous terms. The only difference lies in the co-referential terms 'Hesperus' and 'Phosphorus.' But since two assertions that differ in truth value must also differ in content, it seems that the basic propositions meant with (1) and (2) must be different. Hence, this intuition makes belief reports like those typically made with (1) vulnerable to substitution of co-referential expressions inside the 'that'-clause or, for short, opaque.

Now take (C) and the opacity of belief reports made with (1) and (2) as premises. Since, in virtue of the latter, the basic propositions typically meant by a speaker with (1) and (2) differ and the sentences have the same structure then, in virtue of (C), the content thereby meant with at least one component of (1) must differ from the content meant with the corresponding component of (2). And since every component of (1) has the same meaning as the corresponding component of (2), it follows that at least one component of (1) or (2) is typically not used literally.[4] In this chapter, I will follow the lead of this argument and propose a non-literal account of opaque belief reports.

It would be a vain exercise indeed to simply add another account of opaque belief reports to the long list of existing ones. But my bid is bolder. I claim that my account will have all the

[4] This argument is sketched in Szabó 2010, p. 267.

attributes philosophers have in vain wanted from an account of belief reports. The trick is not just to appeal to non-literality but also to better understand what philosophers have wanted from an account of belief reports with the help of the new consensus about linguistic communication forged in the last decade or so.

In section I, I will discuss the standard desiderata of an account of belief reports and argue that the reason why they have often been thought jointly incompatible with opacity is an old-fashioned view of linguistic communication that few experts still embrace. In section II, I will argue that abandoning this view frees us of philosophical preconceptions about the mental and allows us to take our putative judgments about psychological interpretation at face value. In section III, I will propose a metonymic account of how we communicate these putative judgments that vindicates our linguistic introspection. In section IV, I will argue that once the standard desiderata of an account of belief reports are conservatively updated to reflect a more realistic picture of linguistic communication, the proposed account satisfies them all.

I. The Incompatibility Argument

All accounts that uphold the Fregean intuition about the opacity of some belief reports violate either (C) or the following independently plausible principles:

(DR) Direct Reference: Referential expressions such as names contribute only an object to the basic proposition meant with sentences in which they occur.

(SI) Semantic Innocence: 'that'-clauses have the same function in belief reports as in any other linguistic contexts in which they occur (e.g. causal statements, explanations).

The tension existing between the Fregean intuitions and these three principles is captured by the

following argument implicit in much of the debate over belief reports. Since 'that'-clauses contribute a proposition to the content of causal statements and explanations, in virtue of the principle of semantic innocence, so must they to the basic proposition meant with (1) and (2); and since, in virtue of the principle of direct reference, 'Hesperus' and 'Phosphorus' contribute only Venus to the basic proposition meant with (1) and (2), which contains[5] the proposition contributed by the 'that'-clause, then 'Hesperus' and 'Phosphorus' must contribute Venus to this proposition as well. Finally, since (1) and (2) are otherwise identical in structure and components, then, in virtue of the principle of compositionality, the same basic proposition must be meant with (1) and (2). But, then, the belief reports made with (1) and (2) must have the same truth value, contrary to the Fregean intuition. Call it *the incompatibility argument*.

All attempts to uphold the Fregean intuition have accordingly denied (C), (DR) or (SI). Frege denied (SI) by claiming that inside the scope of 'believes' 'that'-clauses behave abnormally, expressions occurring in them referring to abstract entities he called *senses*. More recently, the so-called hidden-indexical theory, advocated by Mark Crimmins and John Perry (1989, Crimmins 1992), has sacrificed (C) by positing unarticulated constituents in the basic proposition meant with belief-reporting sentences.[6] Obviously, this account treats belief reports as cases of underarticulation. Others have sacrificed (SI) by having 'that'-clauses contribute ordered pairs of sub-propositional contents and *ways of thinking* of those contents (Recanati 1993, Ch. 18), yet others by making the verb 'to believe' context-sensitive and the function of the 'that'-clause vary accordingly (Richard 1990).

[5] I mean to be neutral with respect to whether the proposition contributed by the 'that'-clause is referred to or described. See King 2002 for discussion.

[6] However, I doubt that this conclusion follows, mutatis mutandis, from weaker principles of compositionality such as that of *semantic compositionality* (SC) discussed in section IV below, as some have claimed (e.g. Bach 1997, p. 218). Crimmins' talk of articulated constituents being "expressed" in a sentence (1992, p. 10) suggests that they do not correspond to any component of the syntactic structure of the sentence. But I suppose it is open to the hidden-indexical theorist to adopt a more flexible notion of a syntactic component along the lines of Stanley's notion of *logical form* (2000). By the lights of Stanley's account, Crimmins' unarticulated constituents would correspond to a component of the logical form of the sentence and so the hidden-indexical theory would plausibly not violate this principle of compositionality—but, alas, at the cost of violating (SI).

Bach (1997) claims that his account is the exception. Bach hypothesizes that beliefs are ineffable, that one can at best loosely describe them. Consequently, it would be true that different basic propositions are normally meant with (1) and (2) as these can be used to describe different beliefs, but what exactly those beliefs are would not be thereby specified. Bach does not provide details, but if belief reports are descriptions of some kind then belief-reporting sentences are semantically incomplete, for they do not in general involve quantifiers.[7] It would seem, therefore, that Bach's account treats belief reports as cases of underarticulation, much as the hidden-indexical theory does. Obviously, then, the principle of compositionality Bach takes credit for cannot be (C), for it must permit sub-propositional contents to apply to Bach's account. However, conforming to such a principle is an empty victory since the Fregean intuitions are specifically about the truth values of belief reports and so about propositional contents. Consequently, that such a principle is respected by Bach's account has no bearing on the problem of opacity and, in fact, it is also respected by the hidden-indexical theory—contrary to Bach's claim to exceptionality. In section IV, I will further discuss this principle of compositionality and argue that it is nearly trivial in the current state of the debate. For now what matters is that Bach's account is no exception because it violates (C).

Worse, Bach's account is obscurantist as it makes beliefs more mysterious than they seem. He himself admits as much (though I doubt he would put it thus) when he recognizes that his account entails that beliefs are not relations between an agent and a proposition, but instead are ineffable. But the relational view of beliefs is consistent with much linguistic evidence and helps to frame what is otherwise a heterogeneous, multidisciplinary debate. I think it should not be given up easily. Consequently, even though I think the motivation for my proposal quite strong already, I will add another very plausible desideratum to the three principles previously

[7] See Neale 1990 for the standard treatment of descriptions.

stated, namely,

(DAB) The Dyadic Analysis of Belief: All beliefs can be specified by a belief report made with a sentence of the form 'A believes that S' where A is a designating expression and S a sentence.

No attempt to uphold the Fregean intuitions comes close to conforming to (C), (DR), (SI) and (DAB). My proposal will.

Discouraged by the apparent force of the incompatibility argument and unwilling to give up any of the above principles, some have opted for discrediting the Fregean intuitions instead (e.g. Soames 1985, Salmon 1986, Saul 1993, Braun 1998). Even proponents of these accounts would admit that the Fregean intuitions are a desideratum of a theory of belief reports. They sacrifice it because they are convinced by the incompatibility argument that they must choose between the Fregean intuitions and the principles. My goal is also to prove that this is a false dilemma by disarming the incompatibility argument.

I think that the reason why the incompatibility argument has seemed so formidable is an old-fashioned and ultimately inadequate conception of linguistic communication.[8] The last two decades have witnessed the emergence of a consensus that the traditional semantics/pragmatics distinction does not square with our linguistic practices. In particular, it has become clear that actual linguistic communication is typically not divided into a purely semantic stage, when the basic proposition associated with a use of a sentence is decoded from the sentence independently of the speaker's specific communicative intentions; and a purely pragmatic stage, when that proposition is taken as grounds for deriving the speaker's specifically intended message. Instead, it is widely agreed now that the basic proposition associated with the use of a sentence—namely,

[8] Bave (2008) is one of the few working in the area who have not taken for granted the old-fashioned view.

that which the speaker means with the use but does not intend the audience to infer from any other proposition she means with the same use[9]—typically does depend crucially on the speaker's specific communicative intentions. The previously discussed phenomenon of underarticulation has been at the center of these developments. The notions of semantics and pragmatics have been left in need of serious revision or, perhaps, disposal.

The incompatibility argument fails because it assumes that the principles of compositionality, direct reference and semantic innocence are specifically semantic and, therefore, exclude non-literality. For instance, in the above rendition the argument assumes that the principle of direct reference requires that 'Hesperus' and 'Phosphorus' contribute *their* referent (viz. Venus) to the basic proposition meant in belief reports. This is how the principle has traditionally been conceived; but this is a byproduct of the traditional view of linguistic communication. In section IV, I will argue that the principle should be updated to retain its core idea in light of the new consensus. The most natural reformulation of the principle, I will argue, should require something slightly weaker, namely, that referring expressions contribute *an object* to the basic proposition meant with a sentence in which they occur (hence, (DR) above).

What follows turns heavily on the new consensus about linguistic communication. In order not to beg any questions against these developments, I have formulated (C), (DR), (SI) and (DAB) in terms of 'the basic proposition meant with a sentence' as opposed to 'the sentence's meaning' or other tendentious labels. The discussion of section III will specifically justify this usage for opaque belief reports. Moreover, I will defend these formulations of the principles in section IV. If one assumes the traditional semantics/pragmatics distinction, hence the willy-nilly

[9] Sometimes the basic proposition associated with the use of a sentence will be the conjunction of two or more propositions. For example, the basic proposition of normal uses of 'it is raining here in Reykjavik' is *that it is raining in Reykjavik* and *that the speaker is in Reykjavik* (Neale 2007). Furthermore, I restrict all the discussion to the prosaic-assertoric function of language. I specifically exclude poetic-metaphoric uses of sentences in which many optional propositions can be basically meant (e.g. Wilson & Sperber 2002) voiding thereby the uniqueness condition implicit in "the basic proposition."

identification between what *a speaker* basically means with an expression and *the expression's* meaning, (C), (DR), (SI) and (DAB) boil down to the traditional formulations.

II. Ways of Thinking and Folk Psychology

Virtually all proposed explanations of the Fregean intuitions about the truth values of opaque belief reports have a common denominator: they posit a non-intentional determination of mental entities.[10] One of the most widely acquiesced theoretical moves in the face of mind-related riddles is to chalk them up to an extra determination of beliefs, desires, concepts, etc. It is natural to view the intentional content of mental entities as *what* the agent thinks about, and when trying to articulate the extra determination of mental entities, the contrast suggests itself with *how* the agent thinks about it. *Ways of thinking* of contents thus emerge as natural theoretical postulates. Sometimes this extra determination is called *modes of presentations* of contents and sometimes the mental entity's *narrow content*. Sometimes peculiar properties are ascribed to it such as being abstract or being realized in a language-like medium. But never mind what exactly it is, I will use 'ways of thinking' to refer to this alleged extra determination of mental entities in general.

The ways-of-thinking conception of the mind and the traditional semantics/pragmatics distinction form something of a package deal. The two-fold determination of beliefs has a natural fit in the traditional two-stage view of linguistic communication long assumed to govern belief reports. On the one hand, the information encoded by a sentence was often said to be mandatory and characteristically truth-conditional,[11] so it was natural to identify the information encoded by a belief-reporting sentence (or a part thereof) with the intentional content of the belief reported.

[10] The exception is, perhaps unsurprisingly, Bach 1997.
[11] Arguably because such information was thought to be governed by *truth-conditional semantics*.

On the other hand, the information implicated by the speaker was often said to be optional and somewhat secondary to the linguistic act, so it was natural to identify the information implicated with a belief-reporting sentence with the subsidiary way of thinking of the belief's content.

Ways of thinking remain part of the philosophical lore.[12] Yet arguably this place is sustained not by independent evidence but by their perceived need to sort out various puzzles about the mind, not least of which is that of opaque belief reports. It is not my purpose here to mount a full attack on ways of thinking. Theoretical reasons may, on balance, justify clinging to them; but because I am interested in our pre-theoretic intuitions about belief reports, I am interested specifically in the role ways of thinking play in folk psychology. As it turns out, there is very little direct evidence of them playing any role.

For instance, consider what intuitively it takes to understand why Caesar crossed the Rubicon with his army.[13] Asked to explain Caesar's action folks say: "well, he could either go back to Rome leaving his army behind and risking being executed upon his return, or bring the army with him and start a civil war. He thought the second option better." This garden-variety piece of psychological interpretation works by mentioning the options that Caesar considered, relevant consequences of these options, and Caesar's relative evaluation of these pairs. But there is no sign of ways of thinking about these options or their consequences.

How about cases in which the protagonist suffers from an identity error, such as Hammurabi who did not know that the morning star and evening star were one and the same? No reliance on ways of thinking is apparent here either. When filled in on the situation and asked why Hammurabi, say, prayed to the morning star but not to the evening star, folks answer that he wrongly thought that there were two different planets. They do not say that he took Venus in two

[12] Even most Russellians appeal to ways of thinking in explaining why we have Fregean intuitions (e.g. Salmon 1986, Braun 2002).

[13] Although I illustrate the point with practical reason, I believe similar points apply to theoretical reason.

different ways or that he did not identify two ways of thinking about Venus. In fact, if the interpretation of Hammurabi is anything like the interpretation of Caesar, one need only ensure that praying to the morning star and praying to the evening star are distinct options in his deliberation to render him intelligible. This is achieved by bringing ways of thinking to differentiate between options with identical contents, as most philosophers have assumed. But it is equally well achieved by assigning the options different contents altogether, as seems to be implicit in folk explanations, provided that we keep in mind Hammurabi's identity error and so that beliefs assigned different contents can control behavior toward the same object. In sum, ways of thinking are theoretical postulates of which there is no clear trace or need in our folk explanations of agents.

I propose to take the folk explanations at face value. Folk explanations of Hammurabi have it that he believes that there are two objects where in reality there is only one. So I am ready to attribute to Hammurabi beliefs in two distinct objects associated with Venus. Given the transitivity of identity, these two objects cannot be both identical to Venus, for then they would be one and the same. Any object except Venus can play the role of being distinct from Venus. But any object already part of the agent's worldview is automatically disqualified for the role because the gain in interpretative power thus obtained would be offset by an equal loss. Worse, it would be arbitrary to assign any existing object. Hence, I propose that some of Hammurabi's beliefs (among them that reported with either (1) or (2)) are assigned a merely possible object as content. Obviously, this assignment of content is auxiliary, forced by Hammurabi's identity error. Accordingly, it does not bar these beliefs from partially controlling Hammurabi's behavior toward Venus, for it is Venus which occupies in the actual world the role of the merely possible

object in the resulting model of Hammurabi's worldview.[14]

My goal in this chapter is to develop an account of the alleged difference in truth value between belief reports normally made with (1) and (2) as simply due to a difference in the reported beliefs' intentional contents which is conveyed with non-literal uses of 'Hesperus' or 'Phosphorus.'

Which of the two uses of co-referential expressions is non-literal generally depends on the context and the explanatory goals of the parties in the conversation. For example, whether it is 'Hesperus' or 'Phosphorus' which is used to refer to the merely possible object in reporting beliefs with (1) and (2) may depend on which contents had been previously assigned or which belief we see as less central to the range of behaviors we are interested in.[15] Worse, the same belief that on one occasion is assigned the merely possible object as content can, on another occasion, be assigned Venus. For example, when discussing Hammurabi's sleeping patterns, somebody might say: "Oftentimes Hammurabi got up from bed early because he knew that Venus rises in the morning." Not only does this sound harmless but, in the context, it may well be the correct explanation of why Hammurabi got up early at certain times of the year. If so, it would seem that

3) Hammurabi knew that Venus rises in the morning

is true. But then 'Venus' must be interpreted as referring to Venus, as opposed to a merely possible object, for the content that a merely possible planet rises in the morning could not constitute knowledge. But, then, from (3) plus standard analyses of knowledge we can infer:

[14] Though I won't elaborate the point here, I think it is plausible that people do exactly the same when interpreting agents with empty beliefs, e.g. children who believe in Santa.

[15] However, once one is taken non-literally (e.g. 'Phosphorus'), it will be natural in most contexts to take the other one literally (e.g. 'Hesperus' in (2) to refer to Venus itself) if only for considerations of charity (see below). I will assume so for the sake of simplicity.

4) Hammurabi believed that Venus rises in the morning

where 'Venus' must contribute Venus itself to the intentional content of the belief reported. I claim that this is the same belief as that reported with (1) in a context in which 'Phosphorus' is used to refer to a merely possible object. In sum, the proposal requires that assignments of content to mental states may vary with context and explanatory goals of the interpreters.

Some may be troubled by this postulated indeterminacy of content and consider it a *reductio* of my attempt to dispense with ways of thinking. But this indeterminacy is no *ad hoc* stipulation: cases of identity error notoriously induce indecision between alternatives in folk attributions of content. Moreover, ways of thinking are utterly powerless to discipline some of these cases.

Consider cases of error through misidentification (e.g. Pryor 1999). Imagine, for example, that you take the elevator at your ophthalmologist's building and, out of sheer luck, recognize Saul Kripke wearing a red shirt as the only other occupant. The doors open before reaching the ground floor and Kripke walks out while someone who looks like the building's janitor walks in. In passing, however, the man says a few things to Kripke that he obviously fails to understand. Before Kripke has time to explain to the man that he has confused him with another person, the doors close. You are left in the elevator with the man who is in error about Kripke's identity. Does the man believe that Kripke is wearing red shirt? It depends. If the question is asked by a color-blind specialist, the answer is presumably *yes*—he did form the belief that Kripke is wearing a red shirt. But if the question is asked by a philosopher, the answer seems to be *no*—the man does not know who Kripke is and so hardly can he believe that Kripke is wearing a red shirt. I don't think either answer is more or less correct than the other independently of context and our explanatory goals. Rather, there is just no absolute fact as to

what is the content of the man's mental states modulating his behavior toward Kripke. Moreover, ways of thinking cannot come to the rescue here because the problem with this man is that his thought is more coarse-grained than the world!

This is not to say that it is always indeterminate what the content of our mental states is or that when it is indeterminate, it is entirely arbitrary what it is. Rather, I claim that under specific and relatively rare circumstances (viz. identity errors) it may be indeterminate what the content of an agent's mental states is and that this indeterminacy is limited to a few possibilities. In much the same way, many (perhaps most) of our concepts become indeterminate under special circumstances. Consider our concept of table. There are scores of cases where it seems absolutely determinate whether something is or is not a table. But there also are borderline cases in which whether an object qualifies as a table seems to boil down to how it is used and so depends on our interest in the practices of those who use it.

There is a long tradition of philosophers who have allowed for moderate indeterminacy in the content of thought, including Quine (1960), Davidson (1973), Lewis (1994) and Williamson (2008, Ch. 8). Lewis, for example, claimed that the content of someone's mental states is determined by the assignments of referents that maximize true belief[16]—what he calls *the principle of charity*. But, in cases of identity errors such as those described, there may not be an assignment of referents which maximizes true belief because different assignments can make incompatible yet equally numerous beliefs true. In such cases, which assignment are more *charitable* may depend on the salience in the context of some aspects of the agent's epistemic make-up. For instance, focusing on Hammurabi's identity error precipitates assignments of referents to his beliefs which avert the attribution of a contradiction; hence the intuitive difference in truth value between uses of (1) and (2). But when the identity error is marginal in

[16] Given a restriction to natural properties (see Lewis 1984, 1994).

the context (e.g. Hammurabi's sleeping patterns) true belief seems maximized by assigning Venus as the content of the relevant beliefs.

I will hereafter assume that the interpretation of agents like Hammurabi involves some indeterminacy of content rooted in their identity error. Hence, when I propose an account of the basic proposition meant in belief reports such as those typically made with (1) or (2), I will be assuming that the agent's identity error is salient in the context and that the speaker's goal is, for whatever reasons, to emphasize the error's effect on the interpretation of the agent.

III. Metonymy

Sometimes called *deferred reference* (Stanley 2005), *semantic transfer* (Recanati 2004), or *ad hoc concept construction* (Carston 2002a), metonymy is a linguistic phenomenon in which a speaker relies on the contextual contiguity between two objects to designate one with an expression whose semantic value is the other.[17] For example, someone might say

5) Washington DC is corrupt

to indicate that the politicians who live in Washington DC are corrupt. The speaker relies on the fact that corruption is a property of people only to secure reference to the some people characteristically associated with Washington DC, not to the city itself. Metonymy is akin to nicknaming and so obviously pervasive in everyday linguistic communication. So much so that it is thought to play a crucial role in the evolution of natural languages (Carston 2002a).

Metonymy is arguably the linguistic phenomenon that most clearly blurs the line between semantics and pragmatics as traditionally understood. Even those who are skeptical of other

[17] I deliberately define metonymy in a way that sweeps in some metaphoric uses of designating expressions since one type of *contextual contiguity* is similarity. This is in part because I doubt that metonymy and metaphor can be defined exclusively. But, most importantly, even if there is a more restrictive definition of metonymy, I want my proposal to be ecumenical.

alleged pragmatic intrusions into the semantic province—such as the previously mentioned phenomenon of underarticulation (e.g. Stanley 2000, Stanley & Szabó 2003)—concede that metonymy defies categorization along the traditional lines (Stanley 2005). The reason is that, on the one hand, by definition an object designated metonymically departs from the information encoded by the expression used and so hardly can metonymy be categorized as a semantic phenomenon. But, on the other hand, metonymic uses are typically not subordinate to the proposition encoded by the sentence used. For example, in uttering (5), the speaker above clearly does not intend the audience to consider the proposition having the city of Washington as a constituent and then use it as a premise wherefrom to infer the proposition having the politicians who reside in the city as a constituent. Rather, the speech act is planned so that the latter proposition is constructed directly from the constituents picked out metonymically.[18] This makes metonymy unsuited for the traditional, Gricean conception of pragmatics.

Metonymy exemplifies the intrusion of the speaker's specific communicative intentions in the construction of the basic proposition from a use of a sentence. This is important for my account because we are not aware of relying on conversational implicatures when making or processing opaque belief reports, as some insist we do (Salmon 1986, Saul 1998). If we are to trust our linguistic introspection, since the Fregean intuitions are about the truth value of uses of sentences (hence, about their propositional content), the reporter's specific communicative intentions must somehow make their way to the basic proposition she means. Metonymy provides an elegant explanation of how this happens.

Following the discussion of the previous section, I will assume that when people report others' beliefs they are primarily providing grist for psychological interpretation.[19] I will also

[18] For a dissenting opinion, or otherwise doubts, see Bach 1994b.

[19] I will focus on others' beliefs because oftentimes when people report their own beliefs they are primarily making assertions

assume that when people interpret an agent suffering from an identity error crucial to her intelligibility, they do so by attributing beliefs about a merely possible object. In a nutshell, my proposal is that when reporting beliefs critically associated with identity errors people as usual describe a dyadic relation between the agent and a proposition, but they pick out the merely possible constituent of that proposition metonymically. For example, with (1), depending on the context, I may want to communicate that Hammurabi believes that a certain non-existing planet rises in the morning. To do so, I conspicuously use the word 'Phosphorus' which refers to Venus but, because I take the audience to be aware of Hammurabi's identity error and its relevance to the interpretative situation, will get the message across much in the way 'Washington DC' does in (5). Obviously, my professed goal in doing so is to draw a contrast with the report I would make with (2) in the same context and so the basic propositions typically meant with (1) and (2) differ in truth value. This both provides a folk-psychological basis for the Fregean intuition and confirms our linguistic introspection.

Whether implicated or not, any significant departure from semantic conventions postulated to account for a linguistic phenomenon must be justified by a practical gain to be plausible. The exact form of this justification is the subject of some controversy but here I will consider the elegant and widely influential account known as *relevance theory* (Sperber & Wilson 1986, Carston 2002a).[20]

According to this account, every ostensive act of communication carries, and is based on, the mutual expectation of optimal *relevance*. Relevance can be achieved in two non-exclusive ways: a) by maximizing the cognitive effect of the act of communication in the audience, or b)

---

and/or trying to guide the audience to (what they think is) the truth (e.g. Devitt 1996). Plus, identity errors cannot be self-ascribed, for then one would have all the reason to correct them.

[20] My choice of relevance theory is partly based on the fact that it is supposed to somehow generalize Grice's account. The original formulation of the theory (Sperber & Wilson 1986) has undergone some revisions, but for the sake of simplicity I will ignore these developments. Nothing of importance to my proposal turns on them.

by minimizing the cost of processing the act. Cognitive effects are understood as changes in belief, which can occur by the formation of new beliefs or by the abandonment of old ones. Processing costs, on the other hand, are understood in terms of the accessing of information. Here context plays an important role since information that has already been accessed is thereafter cheaper to access in the same context. Roughly, this determines that, in processing linguistic acts, people will, other things equal, seek interpretations that are less costly, given the context, and more informative. Meanwhile, aware of this, speakers will plan their linguistic acts so as to ensure that the intended interpretation is optimal in this sense. For example, trivial pieces of information will usually be passed over given their null cognitive effect, while easier or readily available interpretations in a context will, other things equal, take priority over extraneous or more convoluted ones.

In the case of belief reports such as those made with (1) and (2), we are assuming that the agent's identity error is already salient in the context and so cheaply accessible. On the other hand, given the discussion of the previous section, attributing to someone an identity error presupposes the attribution of scores of mental states modulating her behavior toward the object of error and, accordingly, the information that the agent has certain beliefs about that object independently of the identity error will have relatively little cognitive effect. Finally, because identity errors are typically idiosyncratic, if agents who suffer from them are in fact interpreted by appeal to a merely possible object, we typically lack names for such an object. As a consequence, it can only be designated semantically by description. However, such a way of specifying the intentional content of the belief is costly and cheaper ways of achieving the same result will be preferred. Given that the audience is aware of the identity error and its relevance to the interpretative situation, the merely possible object is sufficiently contextually contiguous to

Venus to sustain metonymy, making it the natural choice.

Though we may usually lack names for the merely possible objects assigned as contents in opaque belief reports, I am not claiming that all designating expressions are equally well-suited to do the metonymic work. For example, sometimes the agent herself will describe her beliefs by drawing a contrast between the two objects she believes to exist with the help of two or more names. Such is the case of Babylonians who used the name 'Phosphorus' or 'Hesperus' [21] to distinguish between Venus and the extra object they believed to exist. Unbeknown to them, these names were co-referential; but a history of misuses can, too, be exploited in linguistic communication. In fact, I say that the Fregean intuition about reports of the Babylonians' beliefs is typically illustrated with these names because, given the history of their misuse, they are better suited to do the metonymic work than, say, 'Venus.' Put in terms of relevance theory, the merely possible object involved in reports typically made with (1) is readily accessible given the background information about how Babylonians used these names vis-à-vis their identity error.

This is not to say that 'Venus' could not possibly do the job. Kripke (1979) has famously discussed cases of identity error in which speaker and audience lack any special linguistic resource to draw the contrast between the contents of relevant beliefs, yet achieve a strikingly similar communicative effect. Perhaps the most straightforward case illustrated by Kripke is that of Paderewski. Peter learned long ago that Paderewski is a great pianist. Without so much as entertaining that thought again, Peter sees on the news that Paderewski was elected prime minister of Poland. Peter generally thinks that politicians have no musical talent, without realizing that Poland's prime minister is no other than the great pianist. But, then, it seems that I could say truly:

[21] In fact, their corresponding translations.

6) Peter believes that Paderewski has no musical talent.

And it also seems I could say truly:

7) Peter believes that Paderewski has musical talent.

The problem is that jointly (7) and (8) seem to impute irrationality to Peter yet, given his epistemic situation, this is clearly incorrect. My proposal avoids the imputation of irrationality by treating (7) and (8) as conveying opaque belief reports. What happens in Kripke's examples, according to my proposal, is that recognizing the lack of other means of securing metonymy, the audience is ready to go along with the least conspicuous of names.

But also there are circumstances in which descriptions of the merely possible constituent of the proposition believed may be necessary despite their higher processing costs. One such example is conjunctive sentences of contrastive beliefs based on identity errors where no names are available to draw the contrast; for example,

8) Peter believes that Paderewski has no musical talent but he does not believe that Paderewski has no musical talent.

In cases like this, the basic linguistic assumption that expressions are used uniformly within a sentence raises the probability of misunderstanding beyond what may be tolerable in the context. This is not to say that the audience could not possibly get the message but only that the risk of misunderstanding may be too high for the speaker to take. In fact, when such cases are discussed in the literature of belief reports, alas, descriptions are recruited for service:

9) Peter believes that Paderewski, *the prime minister*, has no musical talent but he does not believe that Paderewski, *the pianist*, has no musical talent (Bach 1997, p. 224).

I conclude by arguing that the proposal satisfies the standard desiderata of an account of belief reports.

IV. Desiderata under the New Consensus

Direct Reference:

Traditionally, the principle of direct reference has been formulated thus:

(TDR) Direct Reference: Singular terms contribute only their referents to the proposition semantically expressed by a sentence in which they occur.[22]

However, this formulation is a remnant of the old-fashioned conception of linguistic communication implicit in the traditional semantics/pragmatics distinction. Because on this conception the basic proposition meant by *a speaker* with a sentence is willy-nilly identified with the proposition semantically expressed by *the sentence*, which is determined independently of the speaker's specific communicative intentions, a name's contribution to the basic proposition meant has been thought to be, invariably, its referent or semantic value (e.g. Kripke 1980, Kaplan 1977/1989, Soames 1989). However, when we abandon the old-fashioned view, this rationale breaks apart with the break between the proposition semantically expressed by a sentence and the basic proposition meant with it. Cases of metonymy show that a name can contribute to the latter not its referent or semantic value but an otherwise unrelated object fixed by the speaker's specific communicative intentions.

To defend (TDR) one might dig in one's heels and restrict its application only to the proposition semantically expressed by a sentence even if this has nothing to do with the basic proposition meant. Thus, one might insist that, though it may be irrelevant to the basic

[22] See, for example, Bach 1997.

proposition meant with (5), it is still true that 'Washington DC' contributes only the Washington DC to the proposition semantically expressed. However, this modesty would betray the spirit of the principle of direct reference.

The principle was introduced as a challenge to the then dominant view that the function of names in linguistic communication is reducible to that of descriptions (Kripke 1980). The treatment of names prescribed instead was modeled after the linguistic function of indexicals (Kaplan 1977/1989). Today it is widely accepted that many (perhaps most) uses of indexicals (e.g. demonstratives) rely on the specific communicative intentions of the speaker. So, if the cases after which this theory of names was modeled already involve non-semantic elements, the semantic restriction cannot be an essential ingredient of the picture underlying the theory. Plausibly, the underlying picture is about linguistic communication at large and it turns on the idea that the function of language is primarily to serve as a cue to coordinate the attention of speaker and audience on certain objects, rather than to synchronize their thoughts in the abstract. If so, cases of metonymy only strengthen the challenge to the description theory because, even if names were analyzable as descriptions, their purported function of providing identification conditions of the name's contribution to the basic proposition meant would be ineffectual in metonymic uses. Worse, descriptions themselves can be used metonymically in a way that renders their descriptive content largely irrelevant to what the speaker says, as in (5).

In sum, rather than restricting the principle to exclude metonymy and other linguistic phenomena that trouble the old-fashioned view of linguistic communication, we should reformulate it to incorporate these phenomena in order to preserve its spirit. (DR), for which my proposal was custom-made, does exactly that.

Semantic Innocence:

It might be thought that my proposal violates our semantic innocence since metonymy essentially involves a departure from semantic rules. However, this is not a violation of *the principle of semantic innocence* as it features in the debate over belief reports. This principle pertains specifically to the *function* of 'that'-clauses inside belief reports. In particular, the principle is not supposed to exclude the possibility that 'that'-clauses involve non-literal elements.

The confusion stems, yet again, from the old-fashioned view of linguistic communication. If the contribution 'that'-clauses made to the basic proposition meant with a sentence could be confined to their semantic value independently of the speaker's specific communicative intentions, it would perhaps be natural to constrain the principle of semantic innocence to the semantic behavior of 'that'-clauses, leaving out metonymy and its ilk. But since the speaker's specific communicative intentions already infect the interpretation of 'that'-clauses (e.g. demonstrative expressions can occur in them), such a construal is unduly restrictive.

More to the point, if metonymy does not disturb the function of 'that'-clauses in other contexts such as causal statements and explanations, it cannot possibly violate the principle of semantic innocence. As it happens, all other contexts in which 'that'-clauses feature clearly allow for metonymy. For example, in

10) The cause of the decline of the country is that Washington DC is corrupt

'Washington DC' is obviously used metonymically and yet this in no way upsets the normal linguistic function of the 'that'-clause. My proposal does not, therefore, violate the principle of semantic innocence.

On the contrary, my proposal reaffirms the view that the function of 'that'-clauses is to pick out a proposition in all their uses, whether literally or otherwise, and this seems to be the

core idea behind the principle of semantic innocence.

Compositionality:

Given the argument with which I opened this chapter, my proposal is obviously compatible with (C), for it actually follows from it (plus some plausible assumptions). However, when philosophers have discussed compositionality in relation to opaque belief reports they have often had a weaker principle in mind (e.g. Salmon 1986, Ch. 4; Schiffer 1992; Bach 1997), namely,

(SC) Compositionality: The semantic value of a complex expression is determined by its structure and the semantic value of its constituents.

It is not surprising that this has been the principle that has featured most prominently in debates over belief reports, given the old-fashioned view of linguistic communication that has framed them. After all, as we saw in the introduction, under the widely held assumption that opaque belief reports are literal, the incompatibility argument, together with the non-negotiability of (DR) and (SI), ensures that the Fregean intuitions violate (C). But, moreover, when it is also assumed that the basic proposition meant with a (declarative) sentence just is the sentence's semantic value, (C) boils down to (SC) restricted to full sentences. In short, on the traditional terms of the debate, the Fregean intuitions violate (SC) too.

This is partly why Frege's puzzles have remained striking, for otherwise the abandonment of the traditional semantics/pragmatics distinction has relegated (SC) to near triviality. In effect, the new consensus strips the semantic value of sentences of its former distinction in linguistic communication and so the only relevant role left for (SC) to play in it is to govern sub-propositional contents which may or may not be constituents of the basic proposition meant with a sentence (Carston 2002b, Bach 2001). With semantic value on the

sidelines, (SC) boils down to the innocuous claim that in the language in question information is encoded consistently with formation rules.[23] But the principle does not claim any kind of explanatory priority of the semantic value of the most basic expressions; nor that the relevant structure must be the superficial syntactic structure of the expression; nor, in fact, that the information encoded in an expression is consciously accessed in the process of linguistic comprehension or even plays a role in it (e.g. Carston 2002a, Bach 2001, Recanati 2004, etc.). In short, (SC) simply does not intersect with many of our controversial intuitions and theoretical debates. Since (C) entails (SC)[24] and is far from trivial,[25] I take it that I set the bar higher for my account in attempting to conform to it.

I should make clear, however, that (C) avoids triviality only in some of the ways that (SC) earns it. While (C) strongly suggests that that the relevant structure is roughly the superficial syntactic structure (Szabó 2010) and that linguistic comprehension proceeds through the recognition of the contents meant with the component of the expressions used, it does not claim explanatory priority for these contents. This is a point of some importance for a metonymic account since, oftentimes in cases of metonymy, the felicity of the linguistic act performed in using a sentence and the felicity of those performed in thereby using its components are interdependent. The audience will normally seek an adjustment of the contents of the linguistic act made with the sentence and those made with its composing expressions which optimizes relevance while the speaker, aware of this fact, will plan the act accordingly (Carston 2002b). In fact, the interdependence is not only between the whole sentence and its components but also sideways among components (Cf. Cohen 1986). For example, part of why 'Washington DC' is so naturally interpreted metonymically in (5) is that 'is corrupt' is categorically incongruent with

[23] Accordingly, this principle is respected by construction by artificial languages (Szabó 2010).
[24] Restricted to indicative mood, but this includes all the relevant cases here.
[25] See, for example, Travis 1994 and Fodor 2001 for criticism.

its literal interpretation.

The Dyadic Analysis of Belief:

Finally, my proposal does not respect (C), (DR) and (SI) at the cost of making beliefs ineffable and thus more obscure than they seem, as Bach's (1997) does. Rather, my proposal is suitable to the standard view that beliefs are dyadic relations between an agent and a proposition. The proposition is contributed by the 'that'-clause of a belief-reporting sentence and does not include ways of thinking about another proposition or content. Rather, it is just a good old Russellian-style, structured proposition.

It might be objected that Russell's view of propositions might not have allowed for merely possible constituents (1905a), as is required by my proposal.[26] But I do not worry about this because I am not wedded to Russellian dogma. What I call Russellianism about propositions is the framework developed in Soames 1987 and motivated by his critique of truth-conditional accounts of propositions.[27] This framework has risen to prominence on the promise of, inter alia, completing Kripke's agenda (Soames 2002) and so, in particular, of issuing propositions that can be the bearers of alethic properties. Yet it is hard to imagine that accounts built within this framework could fulfill this promise without allowing for at least some merely possible objects as constituents of propositions. Unsurprisingly, Russellians have generally done so (Salmon 1987, 1989; Soames 2002, Ch. 3).[28]

V. Some Gricean Evidence

Although Grice was one of the precursors of the traditional semantics/pragmatics distinction and

[26] I should say, however, that I think that the core of my proposal compatible with other more eccentric treatments of empty naming roughly in the tradition of Russellianism but which do not appeal to merely possible objects (e.g. Sainsbury 2007).

[27] For further developments see King 1994, 1995, 2002.

[28] These accounts do so by claiming that one can be acquainted with a merely possible object. I'm not sure of the intelligibility of this claim but will take a rain check on this issue.

viewed conversational implicatures as the paradigm of contents not determined by conventions governing the expressions used, many of the marks of conversational implicatures he discusses apply, mutatis mutandis, to *implicit* contents in general[29] (e.g. Carston 2002a, p. 138). For example, default interpretations of unarticulated or ambiguous constituents are cancellable, in Grice's sense, and even standard assignments of reference to indexical expression can often be cancelled (Crimmins 1992, Ch. 1). A fortiori, the widely recognized cancellability of the opacity of belief reports (e.g. Richard 1987, Saul 1998) is compatible with the hidden-indexical theory. The cancellability of the opacity of belief reports is also entailed by my proposal but, moreover, I claim that some natural ways of cancelling it strongly suggest that the kind of pragmatic phenomenon at work is metonymy rather than saturation, as the hidden-indexical theory has it.

I will distinguish between two types of cancellability which Grice and much of the subsequent literature have treated as two facets of the same test:[30] basic cancellability and meta-cancellability—as I shall call them. The proposition, P, communicated in uttering an expression, E, is basically cancellable iff an assertion of ~P can be suffixed to an utterance of 'E but' in the same context to perform a non-contradictory assertion. On the other hand, the proposition, P, communicated in uttering an expression, E, is meta-cancellable iff P is explicitly retractable by literal meta-assertion, that is, by an assertion about the speech act one performed in uttering E which does not ascribe non-literality. Examining meta-cancellability has the advantage for our purposes that it forces a more or less specific pronouncement about the type of pragmatic inference involved in arriving at what is communicated with the utterance and will be my focus here. Since the hidden-indexical theory and my proposal postulate different pragmatic processes associated with opaque belief reports, the application of the meta-cancellability test should spur

[29] Including Grice's conventional implicatures.

[30] Cf. "…a putative conversational implicature that *p* is explicitly cancelable if, to the form of words the utterance of which putatively implicates that *p*, it is admissible to add but not *p*, or *I do not mean to imply that p*" (Grice 1967b/1989, p. 44)

our intuitions about which one is correct.

Here are some examples of statements that meta-cancel some natural pragmatic inferences:

11) It is raining—and I am not talking about *here*.

Or (5) could be meta-cancelled thus:

12) The Washington DC is corrupt—and I am talking literally

or, simply,

13) Literally—Washington DC is corrupt.[31]

Also conversational implicatures are generally meta-cancellable as in the following answer to the question whether you are coming to the party:

14) I have to study—and by that I am not implying that I will not go to the party.

According to the hidden-indexical theory, the meta-cancellation statements for (1) would be something like

15) Hammurabi believes that Hesperus rises in the morning—and I am not talking about Hammurabi's way of thinking of Venus as it appears in the morning

while, according to my proposal, it would be something like

16) Hammurabi believes that Hesperus rises in the morning—and I am talking literally.

[31] Of course, circumstances that can mitigate the category incongruence of the literal interpretation are abnormal, but do exist. For example, I could be talking to my wife about what's happening in the cartoon our 1-year-old son is watching in which city are animated characters.

or, simply,

17) Hammurabi believes—literally—that Hesperus rises in the morning.

I argued in section II that folk explanations of agents like Hammurabi do not involve ways of thinking about the object of identity error. I believe this is corroborated by the fact that (17) sounds awkward even to people enlightened regarding the specifics of Hammurabi's identity error. In contrast, the likes of (18) have been used independently to cancel the opacity of belief reports by, alas, advocates of ways of thinking. Ray Jackendoff (1983, p. 216) offers 'in effect' as a transparency-inducing operator while Peter Ludlow (1995b, p. 105), following Jackendoff, suggests 'so to speak' as an opacity-inducing operator. They claim that

18) Hammurabi believes—in effect—that Hesperus rises in the morning

and

19) Hammurabi believes—so to speak—that Hesperus rises in the morning

convey transparent and opaque belief reports, respectively. That these locutions work as meta-cancellation and meta-reinforcement devices, respectively, is suggested by their occurring as subordinate clauses. The most natural interpretation of 'in effect' and 'so to speak' in the above reports is, in fact, as variants of 'literally' and 'non-literally.'

In sum, (19) and (20) have all the appearance to work by disavowing and reinforcing some non-literality built in belief reports normally made with (1). This appearance is compatible with metonymy but not with saturation as the pragmatic inference underlying opaque belief reports. Consequently, linguistic evidence offers some prima facie support for the metonymic account here proposed over the hidden-indexical theory.

VI. Nondetachability and the 'of'-Construction

Finally, I would like to anticipate an objection and, in passing, impugn the evidence for the scope-ambiguity account of opacity (Quine 1956). In addition to the cancellability test, Grice also proposed the nondetachability test for conversational implicatures. As with the first, the reasons that justify the nondetachability test's application to conversational implicatures seem to justify its application, mutatis mutandis, to pragmatic inferences generally.[32] Thus extended, the test says that when a pragmatic inference is not carried by the manner of expression, it should not be possible to find a synonymous expression which does not carry it. So, if the opacity of belief reports can be detached in this way when they do not rely on the manner of expression for their felicity, this would constitute evidence against the hypothesis that opacity results from a pragmatic inference. And, in fact, it seems that opaque belief reports such as (8), which do not rely on the manner of expression, can be turned into transparent ones with the help of Quine's 'of'-construction (1956), namely, as

20) Peter believes *of* Paderewski that he has musical talent.

However, Grice was careful to point out that the nondetachability test admits of exceptions such as when there is no synonymous expression or only "…one which will introduce peculiarities of manner, such as by being artificial or long-winded" (Grice 1989, p. 43). I will argue that (21) is one such exception.

The 'of'-construction ordinarily works as a multipurpose exportation device: it is used any time one wants to predicate something of a proposition and its subject cannot, or need not, be specified as part of the proposition's linguistic expression. For example, when the content of a

[32] Including Grice's conventional implicatures.

belief attribution is unknown or the 'that'-clause cannot, or need not, be specified, or when reference to a proposition is achieved through a demonstrative or name, say, to predicate truth or falsity of it, etc. in all these cases it is natural to resort to the 'of'-construction. Here are some examples:

21) What can you say of Paderewski?

22) That is what I know of Ortcutt

23) Ralph thinks the same as Jones of the man in the brown hat

24) That is true of Hammurabi

25) P is false of most people.

Pace supporters of the scope-ambiguity account of belief reports, it is clear that in at least some of these cases the use of the 'of'-construction does not reveal the underlying syntactic structure of the statement any more than the use of indexical expressions does. Like indexicals, the 'of'-construction is typically used as a device of convenience in expression and, if so, what speakers typically say with it has paraphrases that dispense with it.[33]

Subjunctive sentences involve predication of a proposition and so are suitable to exportation of its subject with the 'of'-construction. This, again, may be a matter of convenience, as in

26) That is necessary of Ortcutt but not of Smith

27) Anything is possible of him

both of which having paraphrases that avoid exportation of the subject of the proposition. However, since quantifiers produce ambiguities when within the scope of a modal operator, the

[33] Provided we tolerate reference to predicates or, in cases like (24), *quantification* over predicates: e.g. Ralph and Jones both think that the man in the brown hat is *something*. See also (28) below.

'of'-construction may also be useful in disambiguating some subjunctive sentences; for example:

28) It is necessary of the number of planets that it is greater than 7.

Here, since the 'of'-construction is mobile, it allows the speaker to signal that the position of the definite description (and quantifier; see Neale 1990) 'the number of planets' is irrelevant to the basic proposition meant with (29) and, in fact, can be exported outside of the 'that'-clause which typically specifies the proposition that is the object of predication. But because the 'of'-construction also has convenience uses, it does not always succeed in disambiguating scope relations. For example,

29) That is necessary of the man in the brown hat

admits of both a scope-disambiguation interpretation and a convenience interpretation. In sum, the 'of'-construction plays a utility role in relation to statements involving predication of a proposition, resolving ambiguities of scope being but one of many possible uses.

But, then, the 'of'-construction is ill-suited to work as a disambiguation construction in favor of the transparent readings of reports of intentional mental states, for some intentional verbs do not involve predication of a proposition but of an object—the so-called *objectual attitudes*. The proposed analysis predicts that the 'of'-construction should be out of its elements in these reports since objectual attitudes do not involve a proposition of which to take the subject out. This is exactly what one observes since

30) Lois Lane loves of Superman

is obviously ill-formed and meaningless, even though

31) Lois Lane loves Superman

is naturally read opaquely. Quite obviously, then, objectual attitudes are as vulnerable to the phenomenon of opacity as its propositional counterparts (e.g. Salmon 1986, Ch. 8). The fact that philosophers have focused on the latter reflects nothing more than a methodological choice and so we should be careful not to conflate what is particular to them with what is essential to the phenomenon of opacity.

In fact, there is an important difference in the use of the 'of'-construction to disambiguate belief reports and subjunctive sentences. Normally refined non-expert users of subjunctive discourse usually get the *de re* reading from utterances of sentences like (29). In contrast, folks react with perplexity to utterances of the likes of (21). Given the view of beliefs as relations between an agent and a proposition, the 'of'-construction is as grammatically acceptable in (21) as in (29). Therefore, the fact that it sounds so artificial in (21) suggests that, unlike in (29), there are no scope issues going on and so that the construction's only role in belief reports is as a convenience device. The artificiality of (21) is thus explained by the futility of this use of the 'of'-construction, there being no need to omit the subject of the proposition designated by the 'that'-clause.

Quine (1956) introduced the 'of'-construction with the professed purpose to express the transparent reading of belief reports. Because of the conspicuous artificiality of its use in belief reports, the construction can be put to mark any linguistic difference observable in them with just a declaration of purposes like Quine's. What one is doing in such cases, however, is to rely on the manner of expression to exclude otherwise available interpretations of the reports—exactly the kind of exceptions to the nondetachability test Grice warned us against. These alternative interpretations may result from pragmatic inferences and so this implicitly technical use of the

'of'-construction does not constitute evidence against my proposal. I claim that within the convention generated by this technical use, (21) should be classed with (18) rather than with (8).

If this conclusion is correct, it should manifest in a spontaneous preference in belief reports for the interpretation of the 'of'-construction as a convenience device over the interpretation as conveying the transparent reading, even among experts, when both interpretations are open. I say that it does. Consider

32) Ralph believes odd things of the man in the brown hat. [34]

Surely, (39) admits of an opaque reading. In fact, against the backdrop of Ralph's story, the opaque reading is nearly unavoidable.[35]

[34] Bach (1994a, p. 198) attempts another counterexample to the 'of'-construction. But I think his success questionable because he, too, relies on the artificiality of the 'of'-construction in some belief reports to introduce a different use. As a result, his use also feels awkward and forced.

[35] Cf. Ralph believes odd things of the man in the brown hat but ordinary things of Ortcutt.

Bibliography

Anscombe, G.E.M. 1958 *Intention*, New York: Cornell University Press.

Aqvist, L. 1974, A new approach to the logical theory of actions and causality, in

Stenlund 1974, pp. 73-91.

Arjo, D. 1996, Sticking Up for Oedipus: Fodor on Intentional Generalizations and Broad Content, *Mind & Language*, 11, 3, 231-245.

Armstrong, D. 1968, *A Materialist Theory of the Mind*, London: Routledge.

Armstrong, D. 1978, *Universals and Scientific Realism*, vols. I and II, Cambridge: Cambridge University Press.

Armstrong, D. 1997, *A World of States of Affairs*, Cambridge: Cambridge University Press.

Armstrong, D. 2004, *Truth and Truthmakers*, Cambridge: Cambridge University Press.

Arpaly, N. 2000, On Acting Rationally against One's Best Judgment, *Ethics*, 110, 488-513.

Aydede, M. & Robbins, P. 2001, *Canadian Journal of Philosophy*, 31, 1, 1-22.

Ayer, A.J. 1952, *Language, Truth and Logic*, New York: Dover Publications, first Dover edition.

Bach, K. 1981, An Analysis of Self-Deception, *Philosophy and Phenomenological Research*, 41: 351–370.

Bach, K. 1994a, *Thought and Reference*, New York: Oxford University Press.

Bach, K. 1994b, Conversational Impliciture, *Mind & Language*, 9, 124-162.

Bach, K. 1997, Do belief reports report beliefs?, *Pacific Philosophical Quarterly*, 78, 215-241.

Bach, K. 2001, You don't say?, *Synthese*, 128, 15-44.

Baier, A. 1971 The Search for Basic Actions, *American Philosophical Quarterly*, 8, 161-170.

Barnes, A. 1997, *Seeing through Self-Deception*, New York: Cambridge University Press.

Barsalou, L. 1999, Perceptual Symbol Systems, *Behavioral and Brain Sciences*, 22, 577-660.

Barwise, J. and Cooper, R., 1981, "Generalized quantifiers and natural language," *Linguistics and Philosophy*, 4: 159–219.

Barwise, J. & Perry, J. 1983, *Situations and Attitudes*, Cambridge, MA: The MIT Press.

Baumeister, R., Bratslavsky, E., Muraven, M., & Tice, D. 1998, Ego Depletion: Is the Active Self a Limited Resource?, *Journal of Personality and Social Psychology*, 74, 5, 1252-1265.

Bave, A. 2008, A Pragmatic Defense of Millianism, *Philosophical Studies*, 138, 2, 271-289.

Bayart, A. 1959, Quasi Adéquation de la logique modale de second ordre S5 et adéquation de la logique modale de premier ordre S5, *Logique et Analyse*, 6-7, 99-121.

Bechara, A., Tranel, D. & Damasio, H. 2000, Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions, *Brain*, 12, 11, 2189-2202.

Bechara, A., Damasio, H. 2000 & Damasio, A. 2000, Emotion, Decision Making and the Orbitofrontal Cortex, *Cerebral Cortex*, 10, 3, 295-307.

Belnap, Perloff & Xu 2001, *Facing the Future*, Oxford: Oxford University Press.

Bermudez, J. 2000, Self-Deception, Intentions, and Contradictory Beliefs, *Analysis* 60(4): 309–319.

Bermudez, J. 2009, *Decision Theory and Rationality*, NY: Oxford University Press.

Block, N. 1990, Can the Mind Change the World?, in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge: Cambridge University Press, pp. 137-70.

Boghossian, P. 1989,Content and Self-Knowledge, *Philosophical Topics*, 17: 5–26.

Bratman M. 1984, Two Faces of Intentions, *Philosophical Review*, 93, 375-405. Reprinted in Bratman 1987.

Bratman, M. 1987, *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press.

Bratman, M., Israel, D. & Pollack, M. 1988, Plans and Resource-Bounded Practical Reasoning, *Computational Intelligence*, 4, 3, 349-355.

Bratman, M. 2006, What is the Accordion Effect?, *Journal of Ethics*, 10, 5-19.

Braun, D. 1998, Understanding belief reports, *Philosophical Review*, 107, 555-595.

Braun, D. 2002, Cognitive Significance, Attitude Ascriptions, and Ways of Believing Propositions, *Philosophical Studies*, 108, 65-81.

Brentano, F. 1874 (1911/1973), *Psychology from an Empirical Standpoint*, London: Routledge and Kegan Paul.

Burge, T. 1979,Individualism and the Mental, in French, Uehling, and Wettstein (eds.) *Midwest Studies in Philosophy*, IV, Minneapolis: University of Minnesota Press, 73–121.

Burge, T. 1986, Individualism and Psychology, *Philosophical Review*, 95, 3-45.

Byrne, A. 2005, Introspection, *Philosophical Topics*, 33(1), 79–104.

Carnap, R. 1956a, *Meaning and Necessity*, Chicago: University of Chicago Press.

Carnap, R. 1956b, Empiricism, semantics, and ontology, in Carnap 1956a, pp. 203–221.

Carston, R. 2002a, *Thoughts and Utterances*, Oxford: Blackwell.

Carston, R. 2002b, Linguistic Meaning, Communicated Meaning, and Cognitive Pragmatics, *Mind & Language*, 17, 127-148.

Chellas, B. F. 1969, *The logical form of imperatives*, Perry Lane Press, Stanford, CA.

Chisholm, R. M. 1956, *Perceiving: a Philosophical Study*, Cornell University Press.

Churchland, P. 1970,The Logical Character of Action-Explanations, *Philosophical Review*, 79, 214-236.

Cohen, J. 1986, How is Conceptual Innovation Possible?, *Erkenntnis*, 25, 221-238.

Cresswell, M.J. 1967, A Henkin completeness theorem for T, Notre Dame Journal of Formal Logic, 3, 3, 186-190.

Cresswell, M.J. 1975, Identity and Intensional Objects, *Philosophia*, 5, 47-68.

Cresswell, M.J. 1991, In Defense of the Barcan Formula, *Logique et Analyse*, 135-136, 271-282.

Cresswell, M.J. 2006, From Modal Discourse to Possible Worlds, *Studia Logica*, 82, 307-327.

Crimmins, M. and Perry, J. 1989, The prince and the phone booth: Reporting puzzling beliefs, *Journal of Philosophy*, 86, 685-711.

Crimmins, M. 1992, *Talk about Beliefs*, Cambridge, MA: MIT Press.

Damasio, A. 1994, *Descartes' Error: Emotion, Reason, and the Human Brain*, New York: G.P. Putnam's Sons.

Dancy, J. 1997, *Reading Parfit*, Blackwell.

Danto, A. 1965, Basic Actions, *American Philosophical Quarterly*, 2, 141-148.

Davidson, D. 1963, Action, Reasons and Causes, *Journal of Philosophy*, 60, 685-699.

Davidson, D. 1970a, Mental Events, in Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory*, London: Duckworth.

Davidson, D. 1970b, How is Weakness of the Will Possible?, in Joel Feinberg (ed.), *Moral Concepts*, Oxford: Oxford University Press.

Davidson, D. 1971, Agency, in Robert Binkley, Bronaugh, R. and Marras, A. (eds.), *Agent, Action, and Reason*, Toronto: University of Toronto Press.

Davidson, D. 1973a, Freedom to Act, reprinted in *Essays on Actions and Events*, Oxford: Oxford University Press, 1980, 63–81.

Davidson, D. 1973b, Radical Interpretation, *Dialectica*, 27, 314–28.

Davidson, D. 2006, *The Essential Davidson*, Oxford: Oxford University Press.

Davies, M. 1981, *Meaning, Quantification, Necessity*. London: Routledge and Keegan Paul.

Dennett, D. 1987, *The Intentional Stance*, Cambridge, Mass.: MIT Press.

de Sousa, R. 1987, *The Rationality of Emotion*, Cambridge, MA: MIT Press.

Devitt, M. 1996, *Coming to our senses*, Cambridge University Press.

Donnellan, K. S. 1966, Reference and Definite Descriptions, *Philosophical Review* 77, 281-304.

Dretske, F. 1981, *Knowledge and the Flow of Information*, Oxford: Blackwell.

Dretske, F., 1988, *Explaining Behavior: Reasons in a World of Causes*, Cambridge, MA: MIT Press.

Egan, F. 1995, Computation and Content, *Philosophical Review*, 104(2), 181-203.

Etchemendy, J. 1990, *The Concept of Logical Consequence*, Cambridge, MA: Harvard University Press.

Evans, G. 1982, *The Varieties of Reference*, Oxford: Oxford University Press (ed. J. McDowell).

Fauconnier, G. & Turner, M. 1999, Metonymy and Conceptual Integration, in Phanther & Radden 1999.

Feinberg, J. 1972, *Doing and Deserving*, Princeton: Princeton University Press.

Fitting, M. 2006, FIOL Axiomatized, *Studia Logica*, 84, 1, 1-22.

Fodor, J. 1968, *Psychological Explanation*, New York: Random House.

Fodor, J. 1974, Special Sciences, *Synthese*, 28, 97-115.

Fodor, J. 1975, *The Language of Thought*, New York: Thomas Crowell.

Fodor, J. 1981, Methodological Solipsism Considered as a Research Strategy in Cognitive Science, *Behavioral and Brain Sciences*, 3, 63-73.

Fodor, J. 1983, *The Modularity of Mind*, Cambridge, MA: MIT Press.

Fodor, J. 1986, Individualism and Supervenience, *Aristotelian Society: Supplementary Volume*, SUPP 60, 235-262.

Fodor, J. 1987, *Psychosemantics*, Cambridge, Mass.: Bradford Books.

Fodor, J. 1990, A Theory of Content, in *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, Bradford Book.

Fodor, J. 1991, A modal argument for narrow content, *Journal of Philosophy*, 88, 5-26.

Fodor, J. 1993, *The Elm and the Expert*, Cambridge, Mass.: Bradford Books.

Fodor, J. 2001, Language, thought and compositionality, *Mind & Language*, 16, 1-15.

Forbes, G. 1990, The Indispensability of *Sinn*, *Philosophical Review*, 99, 4, 535-563.

Friston, K. 2005, A theory of cortical responses, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 360 (1456), pp. 815–836.

Friston, K. 2010, The free-energy principle: a unified brain theory?, *Nature Reviews Neuroscience*, 11, 127-138.

Frege, G. 1892, Über Sinn und Bedutung, *Zeitschriftfür Philosophie und philosophische Kritik*, 100: 25–50. English translation as 'On Sinn and Bedeutung,' in Beaney, M. 1997, *The Frege Reader*, Oxford: Blackwell.

Friedrich, J. 1993, Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena, *Psychological Review* 100:298–319.

Gallup, G. Jr. 1970, Chimpanzees: Self recognition, *Science*,167 (3914): 86–87.

Garnham, A. 2001, *Mental Models and the Interpretation of Anaphora*, Pennsylvania: Psychology Press.

Gergely, G. & Csibra, G. 2003, Teleological reasoning in infancy: the naïve theory of rational action, *TRENDS in Cognitive Sciences*, 7, 7, 287-292.

Ginet, C. 1990, *On Action*, Cambridge: Cambridge University Press.

Greene, JD., Sommerville, B., Nystrom, L., Darley, J., & Cohen, J. 2001, An fMRI Investigation of Emotional Engagement in Moral Judgment, *Science*, 293, 5537, 2105-2108.

Grice, P. 1975, Logic and Conversation, in *The Logic of Grammar*, Davidson, D. and Harman, G. (eds.), Encino, CA: Dickenson, 64-75.

Grice, P. 1989, *Studies in the Way of Words*, Cambridge, MA: Harvard University Press.

Grush, R. 1995, *Emulation and Cognition*, Ph.D. Dissertation.

Harman, G., 1986, *Change in View*, Cambridge, MA: MIT Press.

Heil, J. and Mele, A. (Ed.) 1993, *Mental Causation*, New York: Oxford University Press.

Henkin, L. 1949, The Completeness of the First-Order Functional Calculus, *The Journal of Symbolic Logic*, 14, 159-165.

Hilpinen, R. (Ed.) 1981 , *New Studies in Deontic Logic*, D. Reidel Publishing Company, Dordretch.

Hintikka, J. 1962, *Knowledge and Belief: An Introduction to the Logic of the Two Notions*, Cornell: Cornell University Press.

Hirsch, E. 2002, Quantifier Variance and Realism, *Philosophical Issues*, 12 (Realism and Relativism), 51-73.

Hogarth, R. 1980, *Judgment and choice: the psychology of decision*, New York: Wiley Publishers.

Holton, R. 1999, Intention and Weakness of Will, *Journal of Philosophy*, 96, 241-262.

Hornsby, J. 1993, Agency and Causal Explanation, in Heil and Mele 1993.

Horwich, P., 1990, *Truth*, Oxford: Oxford University Press.

Huemer, M. 2001, *Skepticism and the Veil of Perception*, Lanham, Md.: Rowman & Littlefield.

Jackendoff, R. 1983, *Semantics and Cognition*, Cambridge: MIT University Press.

Jeffrey, R. 1983, *The Logic of Decision*, Chicago: University of Chicago Press.

Johnson-Laird, P. 1983, *Mental Models: Towards a cognitive science of language, inference, and consciousness*, Cambridge: Cambridge University Press.

Johnson-Laird, P., Byrne, R. & Schaeken, W. 2004, Propositional reasoning by model, *Psychological Review*, 99, 3, 418-439.

Johnson-Laird, P. N., Girotto, V. & Legrenzi, P. 2004, Reasoning From Inconsistency to Consistency, *Psychological Review*, 111, 3, 640-661.

Johnston, M. 1988, Self-Deception and the Nature of Mind, in *Perspectives on Self-Deception*, B. McLaughlin and A. O. Rorty (eds.), Berkeley: University of California Press.

Joyce, J. 1999, *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press.

Kahneman, D. 2011, *Thinking, Fast and Slow*, New York: Farras, Strous and Giroux.

Kaplan, D. 1969, Quantifying in, *Synthese*, 19, 178-214.

Kaplan, D. 1977/1989, Demonstratives, in Almog, Perry & Wettstein (eds.), *Themes from Kaplan*, (New York: Oxford University Press), 481-504.

Kim, J. 1984, Epiphenomenal and Supervenient Causation, *Midwest Studies in Philosophy*, 9, 257-270.

King, J. 1994, Can Propositions be Naturalistically Acceptable?, *Midwest Studies in Philosophy*, 19, 53-75.

King, J. 1995, Structured Propositions and Complex Predicates, *Nous*, 29, 4, 516-535.

King, J. 2002, Designating Propositions, *Philosophical Review*, 111, 3, 341-371.

Knobe, J. 2006, The Concept of Intentional Action, *Philosophical Studies*, 130, 203-231.

Knobe, J. 2010, Action Trees and Moral Judgment, *Topics in Cognitive Science*, 2, 555-578.

Korsgaard, C. 1990, *The Standpoint of Practical Reason*, New York: Garland Press.

Kosslyn, S.M. 1980, *Image and Mind*, Cambridge, MA: Harvard University Press.

Kosslyn, S.M. 1994, *Image and Brain: The Resolution of the Imagery Debate*, Cambridge, MA: MIT Press.

Kripke, S. 1959, A Completeness Theorem in Modal Logic, *Journal of Symbolic Logic*, 24, 1–14.

Kripke, S. 1963, Semantical Considerations on Modal Logic, *Acta Philosophica Fennica*, 16, 83–94.

Kripke, S. 1979, A puzzle about belief, in A. Margalit (ed.) *Meaning and Use*. Dordrecht: Reidel, 239-283.

Kripke, S. 1980, *Naming and Necessity*, Oxford: Basil Blackwell.

Kveraga, K., Ghuman, A.S., & Bar. M. 2007, Top-down predictions in the cognitive brain, *Brain and Cognition*, 65, 145-168.

Ladyman, J. and Ross, D. (with Spurrett, D. and Collier, J.) 2007, *Every Thing Must Go: Metaphysics Naturalised*, Oxford: Oxford University Press.

Lakoff, G. & Johnson, M., 1980, *Metaphors we live by*, University of Chicago Press.

Langaker, R. 2000, *Grammar and Conceptualization*, Berlin and New York: Mouton de Gruyter.

Lepore, E. and Loewer, B. 1987, Mind Matters, *Journal of Philosophy*, 84, 630-642.

Leslie, A., Friedman, O. & German, T. 2004, Core mechanisms in 'theory of mind,' *TRENDS in Cognitive Sciences*, 8, 12, 528-533.

Levin, M. 1979, Explanation and Prediction in Grammar (and Semantics), *Midwest Studies in Philosophy*, 2: 128-137.

Lewis, D. 1966, An Argument for the Identity Theory, *Journal of Philosophy*, 63, 17-25.

Lewis, D. 1969, *Convention*, Cambridge: Harvard University Press.

Lewis, D. 1970a, Causation, *Journal of Philosophy*, 70, 556-567.

Lewis, D. 1970b, Anselm and Actuality, *Noûs*, 4, 175–188.

Lewis, D. 1972, Psychophysical and Theoretical Identifications, *Australasian Journal of Philosophy*, 50, 249-258.

Lewis, D. 1974, Radical Interpretation, *Synthese*, 23, 331-344.

Lewis, D. 1979, Scorekeeping in a Language Game, *Journal of Philosophical Logic*, 8, 339-359.

Lewis, D. 1981b, Causal Decision Theory, *Australasian Journal of Philosophy*, 59: 5–30.

Lewis, D. 1983a, *Philosophical Papers I*, New York: Oxford University Press.

Lewis, D. 1983b, New Work for a Theory of Universals, *Australasian Journal of Philosophy*, 61: 343–77.

Lewis, D. 1984, Putnam's Paradox, *Australasian Journal of Philosophy*, 62, 221–236.

Lewis, D. 1986a, Causal Explanation, in Lewis 1986b.

Lewis, D. 1986b, *Philosophical Papers II*, New York: Oxford University Press.

Lewis, D. 1986c, *On The Plurality of Worlds*, Oxford: Blackwell.

Lewis, D. 1988, Belief as Desire, *Mind* (New Series), 97, 387, 323-332.

Lewis, D. 1994, Reduction of Mind, in Samuel Guttenplan (ed.), *A Companion to Philosophy of Mind*, Oxford: Blackwell Publishers, pp. 412-431.

Lewis, D. 1996a, Belief as Desire II, *Mind*, 105, 303-313.

Lewis, D. 1996b, Elusive Knowledge, *Australasian Journal of Philosophy*, 74, 4, 549–567.

Linsky, B., and Zalta, E. 1994, In Defense of the Simplest Quantified Modal Logic, *Philosophical Perspectives 8: Logic and Language*, J. Tomberlin (ed.), Atascadero: Ridgeview, 431–458.

Loar, B. 1972, Reference and Propositional Attitudes, *Philosophical Review*, 81, 1, 43-62.

Ludlow, P. 1995a, Externalism, self-knowledge, and the prevalence of slow-switching, *Analysis*, 55: 45-49.

Ludlow, P. 1995b, Logical Form and the Hidden-Indexical Theory, *Journal of Philosophy*, 92, 102-107.

Lycan, W. 1994, *Modality and Meaning*, Dordrecht: Kluwer.

Marr, D. 1983, *Vision: A computational Investigation into the Human Representation and Processing of Visual Information*, New York: W. H. Freeman and Company.

McCann, H. 1974, Volition and Basic Actions, *Philosophical Review*, 83, 451-473.

McCann, H. 1991, Settled Objectives and Rational Constraints, *American Philosophical Quarterly*, 28, 25-36.

McGinn, C. 1982, The Structure of Content, in Woodfield 1982

McIntyre, A.2006, What Is Wrong With Weakness of Will?, *Journal of Philosophy*103, 284-311.

Melden, A.I. 1961, *Free Action*, London: Routledge and Kegan Paul.

Mele, A., 1992, *Springs of Action*, Oxford: Oxford University Press.

Mele, A. 2001, *Self-Deception Unmasked*, Princeton: Princeton University Press.

Menzel, C. 1990, Actualism, Ontological Commitment, and Possible Worlds Semantics, *Synthese*, 85, 355–389.

Menzel, C. 1991, The True Modal Logic, *Journal of Philosophical Logic*, 20, 331–374.

Mikhail, J. 2007, Universal Moral Grammar, *Trends in Cognitive Science*, 11, 143-152.

Mintoff, J. 1997, Slote on Rational Dilemmas and Rational Supererogation, *Erkenntnis*, 46, 1, 111-126.

Mongin, P. 2000, Does Optimization Imply Rationality?, *Synthese*, 124, 73-111.

Moran, R. 2001, *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton: Princeton University Press.

Moore, G. E. 1953, *Some Main Problems of Philosophy*, London: George, Allen and Unwin.

Nagel, T. 1974, What is it Like to Be a Bat?, *Philosophical Review*, 83, 435–450

Neale, S. 1990, *Descriptions*, Cambridge: MIT Press Books.

Neale, S. 2007, On Location, in O'Rourke & Washington (eds.), *Situating Semantics: Essays on the Philosophy of John Perry*, Cambridge, MA: MIT Press, 251–393.

Newell, A. & Simon, H. 1972, *Human Problem Solving*, Prentice Hall.

Noë, A. & O'Regan, K. 2001, A sensorimotor account of vision and visual consciousness, *Behavioral and Brain Sciences*, 24, 939-1031.

Noë, A. 2004, *Action and Perception*, Cambridge, MA: MIT Press.

Nozick, R. 1969, Newcomb's Problem and Two Principles of Choice, in Nicholas Rescher, ed., *Essays in Honor of Carl G. Hempel*, pp. 114–146, Dordrecht: Reidel.

Orenstein, A. 1978, *Existence and the particular quantifier*, Temple University Press.

Panther, K. & Radden, G. (eds.) 1999, *Metonymy in Language and Thought*, Amsterdam and Philadelphia: Benjamins.

Panther, K. & Thornburg, L. (eds.) 2003, *Metonymy and Pragmatic Inferencing*, Amsterdam and Philadelphia: Benjamins.

Papineau, D. 1993, *Philosophical Naturalism*, Oxford: Blackwell.

Parfit, D. 1971, Personal Identity, *Philosophical Review* 80, 3–27.

Parfit, D. 1984, *Reasons and Persons*. New York: Oxford University Press.

Parker, A. 2003, *In the Blink of an Eye: How vision sparked the big bang of evolution*, Perseu Publishers/Basic Books.

Peacocke, C. 1993, Externalist Explanations, *Proceedings of the Aristotelian Society*, 93, 203-230.

Perry, J. 1980, Belief and acceptance, *Midwest Studies in Philosophy*, 5, 533-542.

Peters, R. 1958, *The Concept of Motivation*, London: Routledge and Kegan Paul.

Pettit, P. and McDowell, J. (Ed.) 1986, *Subject, Thought and Context*, Oxford: Clarendon Press.

Pettit, P. 1986, Broad-Minded Explanation and Psychology, in Pettit and McDowell 1986.

Pettit, P. 1991, Decision theory and folk psychology, in Bacharach and Hurley (eds.) *Foundations of Decision Theory: Issues and Advances*, Oxford: Basil Blackwell.

Pettit, P. & Smith, M. 1997, Parfit's P, in Dancy1997.

Plantinga, A. 1976, Actualism and Possible Worlds, *Theoria*, 42: 139–60.

Pörn, I. 1974, Some basic concepts of action, in Stenlund 1974, pp. 93-101.

Priest, G. 2005, *Towards Non-Being. The Logic and Metaphysics of Intentionality*, Oxford: Clarendon.

Prinz, J. 2004, *Gut Reactions: a Perceptual Theory of Emotion*, Oxford: Oxford University Press.

Prior, A. 1957, *Time and Modality*, Oxford: Clarendon.

Pryor, J. 1999, Immunity to error through misidentification, *Philosophical Topics*, 26, 271-304.

Putnam, H. 1967, The Nature of Mental States, in Capitan and Merrill (Ed.) *Art, Mind and Religion*, Pittsburgh: University of Pittsburgh Press.

Putnam, H. 1975, The Meaning of 'Meaning,' *Philosophical Papers,2: Mind, Language, and Reality*, Cambridge: Cambridge University Press.

Putnam, H. 1983, *Realism and Reason*, Cambridge: Cambridge University Press.

Putnam, H. 1987, *The Many Faces of Realism*, La Salle, IL: Open Court.

Pylyshyn, Z. 1980, Cognitive Representation and the Process-Architecture Distinction, *Behavioral and Brain Sciences*, 3, 1.

Pylyshyn, Z. 1984, *Computation and Cognition: Toward a Foundation for Cognitive Science*, Cambridge, Mass: Bradford Books/MIT Press.

Quine, W. 1948, On What There Is, *Review of Metaphysics*, 2, 21-38.

Quine, W.1953, *From a Logical Point of View*, Cambridge, Mass.: Harvard University Press

Quine, W. 1956, Quantifiers and propositional attitudes, *Journal of Philosophy*, 53, 177-187.

Quine, W. 1960, *Word and Object*, Cambridge MA: MIT Press.

Quine, W. 1969, *Ontological Relativity and Other Essays*, New York: Columbia University Press

Radden, G. & Kovecses, Z. 1999, Towards a Theory of Metonymy, in Panther & Radden 1999.

Railton, P. 1978, A Deductive-Nomological Model of Probabilistic Explanation, *Philosophy of Science*, 45, 206-226.

Raz, J. 2005, The Myth of Instrumental Rationality, *Journal of Ethics & Social Philosophy*, 1,

1-28.

Recanati, F. 1993, *Direct Reference: From Language to Thought*, Oxford: Blackwell.

Recanati, F. 2004, *Literal meaning*, Cambridge: Cambridge University Press.

Rescher, N. (Ed.) 1966, *The Logic of Decision and Action*, Pittsburgh University Press, Pittsburgh.

Richard, M. 1987, Attitude Ascriptions, Semantics Theory and Pragmatic Evidence, *Proceedings of the Aristotelian Society*, 87, 243-262.

Richard, M. 1990, *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*, Cambridge University Press.

Roth, A. 1999, Reason Explanations of Action, *Philosophy and Phenomenological Research*, 59, 839-874.

Rosch, E. 1978, Principles of Categorization, in E. Rosch & B. Lloyd (eds.), *Cognition and Categorization*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 27–48.

Russell, B. 1905a, Review of: A. Meinong, *UntersuchungenzurGegenstandstheorie und Psychologie*, *Mind*, 14, 530–538. Reprinted in 1973, Douglas Lackey (ed.), *Bertrand Russell. Essays in Analysis*, London: Allen and Unwin, 77–88.

Russell, B. 1905b, On denoting, *Mind*, 14, 479-493.

Russell, B. 1912, *The Problems of Philosophy*, reprinted by Oxford University Press 1997.

Russell, B. 1914, *Our Knowledge of the External World*, London: Allen & Unwin.

Russell, B. 1956, *Logic and Knowledge*, R.C. Marsh (ed.), London: Allen & Unwin.

Sainsbury, R. 2007, *Reference without Referents*, London: Oxford University Press.

Salmon, N. 1986, *Frege's Puzzle*, Cambridge, Massachusetts: MIT Press.

Salmon, N. 1987, Existence, in *Philosophical Perspectives 1*, J. Tomberlin (ed.), Atascadero: Ridgeview Press.

Salmon, N. 1989, Illogical Belief, *Philosophical Perspectives*, 3: Philosophy of Mind and Action Theory, 243-285.

Salmon, N. 1998, Nonexistence, *Nous*, 32, 3, 277-319.

Salmon, W.C. 1971, *Statistical Explanation and Statistical Relevance*, Pittsburgh University Press, Pittsburgh.

Saul, J. 1993, Still an attitude problem, *Linguistics and Philosophy*, 16, 423-435.

Saul, J. 1998, The pragmatics of attitude ascription, *Philosophical Studies*, 92, 363-389.

Schick, F. 1991, *Understanding Action*, Cambridge: Cambridge University Press.

Schiffer, S. 1987a, *Remnants of Meaning*, Cambridge, Mass.: MIT Press.

Schiffer, S. 1987, The 'Fido'-Fido theory of belief, *Philosophical Perspectives* 1, 455-480.

Schiffer, S. 1992, Belief Ascription, *Journal of Philosophy*, 89, 449-521.

Searle, J. 1983, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press.

Sehon, S.R. 1994, Teleology and the Nature of Mental States, *American Philosophical Quarterly*, 31, 63-72.

Sellars, W. 1956, Empiricism and the Philosophy of Mind, *Minnesota Studies in the Philosophy of Science*, vol. I, H. Feigl & M. Scriven (eds.), Minneapolis, MN: University of Minnesota Press, 253–329.

Sen, A. 1970, *Collective Choice and Social Welfare*, San Francisco: Holden-Day; republished Amsterdam: North-Holland, 1979.

Sen, A. 1971, Choice Function and Revealed Preference, *The Review of Economic Studies*, 38, 3, 307-317. Reprinted in *Choice, Welfare and Measurement*, Cambridge: Harvard University Press.

Sen, A. 1997, Maximization and the Act of Choice, *Econometrica*, 65, 4, 745-779.

Setiya, K. 2007, Cognitivism about Instrumental Reason, *Ethics*, 117: 647–673.

Schultz, W. 1998, Predictive reward signal of dopamine neurons, *Journal of neurophysiology*, 80, 1-27.

Simon, H. 1957, *Models of Man*, Wiley Publishers.

Simon, H. 1997, *Models of Bounded Rationality*, Cambridge: MIT University Press.

Slote, M. 1989, *Beyond Optimizing*, Cambridge, MA: Harvard University Press.

Smith, M. 1987, The Humean Theory of Motivation, *Mind*, 96, 36-61.

Soames, S. 1985, Lost innocence, *Linguistics and Philosophy*, 8, 59-72.

Soames, S. 1987, Direct Reference, Propositional Attitudes and Semantic Content, *Philosophical Topics*, 15, 47-87.

Soames, S. 1989, Direct Reference and Propositional Attitudes, in Almog, J. Perry & H. Wettstein (eds.), *Themes from Kaplan*, New York: Oxford University Press., pp.

393–419.

Soames, S. 1999, *Understanding Truth*, Oxford: Oxford University Press.

Soames, S. 2002, *Beyond Rigidity: The Unfinished Semantic Agenda of* Naming and Necessity, New York: Oxford University Press.

Sosa, D. 1996, The import of the puzzle of belief, *Philosophical Review*, 105, 373-434.

Sperber, D. and Wilson, D. 1986, *Relevance: Communication and Cognition*, Oxford: Blackwell.

Stalnaker, R. 1981, *Inquiry*, Cambridge: MIT University Press.

Stanley, J. 2000, Context and Logical Form, *Linguistics and Philosophy*, 23: 391–424.

Stanley, J. & Szabó, Z. 2003, On Quantifier Domain Restriction, *Mind & Language*, 15, 219-261.

Stanley, J. 2005, Semantics in Context, in Preyer & Georg (eds.), Contextualism in Philosophy, Knowledge, Meaning and Truth, *Philosophical Books*, 48, 3.

Stenlund, S. (Ed.) 1974, *Logical theory and semantic analysis: Essays dedicated to Stig K anger on his fiftieth birthday,* D. Reidel Publishing Company, Dordrecht.

Stich, S. 1983, *From Folk Psychology to Cognitive Science: The Case Against Belief,* Cambridge: MIT Press.

Strawson, P.F. 1952, *Introduction to Logical Theory*, London: Methuen.

Sutton, R. & Barto, A. 1998, *Reinforcement Learning: An Introduction*, Cambridge: MIT Press.

Szabó, Z. 2010, The determination of content, *Philosophical Studies*, 148, 2, 253-272.

Tarski, A. 1936, Über den Begriff der logischen Folgerung, *Actes du Congrès international de philosophie scientifique*, Sorbonne, Paris 1935, vol. VII, Logique, Paris: Hermann, p. 1–11. Translated as "On the Concept of Following Logically," *History and Philosophical Logic*, 23 (2002), 155-196.

Travis, C. 1994, On constraints of generality, *Proceedings of Aristotelian Society* (New Series), 44, 165-188.

Trope, Y., & Liberman, A. 1996, Social Hypothesis Testing: Cognitive and Motivational Mechanisms, in E. Higgins and A. Kruglanski (eds.), *Social Psychology: Handbook of Basic Principles*, pp. 239–70.New York: Guilford Press.

Tversky, A. & Kahneman, D. 1974, Judgment under Uncertainty: Heuristics and Biases, *Science*, 185, 4157, 1124-1131.

van Fraassen, B. 1980, *The Scientific Image*, Oxford University Press, Oxford.

Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J. 2001, The brainweb: Phase synchronization and large-scale integration, *Nature Reviews Neuroscience*, 2 (4) (2001), 229–239

von Wright, G.H. 1966, The Logic of Action: A Sketch, in Rescher 1966, 121-136.

von Wright, G.H. 1981, On the Logic of Norms and Actions, in Hilpinen 1981, 3-35.

Williams, B. 1979, Internal and External Reasons, reprinted in *Moral Luck*. Cambridge: Cambridge University Press, 1981: 101-13.

Williamson, T. 1998, Bare Possibilia, *Erkenntnis*, 48, 257–273.

Williamson, T. 2001, The Necessary Framework of Objects, *Topoi*, 19, 2, 201-208.

Williamson, T. 2008, *The Philosophy of Philosophy*, Malden, MA, et al: Blackwell.

Wilson, D. & Sperber, D. 2002, Truthfulness and Relevance, *Mind*, 111, 583-632.

Wilson, George, 1989, *The Intentionality of Human Action*, Stanford, CA: Stanford University Press.

Wilson, R. 1994, Causal Depth, Theoretical Appropriateness and Individualism, in *Psychology, Philosophy of Science*, 61, 55-75.

Wimmer, H. & Perner, J. 1983, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception, *Cognition*, 13, 1, 103–128.

Woodfield, A. (ed.) 1982, *Thought and Object*, Oxford: Oxford University Press.