University of Alberta

Intentionality in Action:
Looking for "Life" in All the Wrong Places

by

Mason Daniel Cash ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy

Department of Philosophy

Edmonton, Alberta

Fall, 2000

0-612-59569-2

University of Alberta

Library Release Form

Name of Author: *Mason D. Cash*

Title of Thesis: *Intentionality in Action: Looking for "Life" in All the Wrong Places*

Degree: *Doctor of Philosophy*

Year this Degree Granted: *2000*

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Department of Philosophy,
University of Alberta
Edmonton, AB
Canada T6G 2E5

Date: *4 August, 2000*

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Intentionality in Action: Looking for Life in All the Wrong Places* submitted by *Mason Cash* in partial fulfillment of the requirements for the degree of *Doctor of Philosophy*.

F. Jeff Pelletier

Wes Co

Wes Cooper

Michael Dawson

Cressida Heyes

Cressida Heyes

Don Kuiken

Jeff Foss

Date: 26 July 2000

Abstract

Here I outline an "embodied action" approach to Cognitive Science, whose central assumption is that human beings are essentially embodied, embedded in a world and situated in a social context. I present a naturalized account of intentionality from this perspective.

I give a normative account of language-use, as the performance of speech acts as moves within shared norm-governed practices. I then show how the normative practice of giving reasons for actions licenses us to attribute intentional states to people as reasons for their actions, and licenses us to expect people to be committed to acting in certain ways, based on the intentional states that they recognize are appropriately attributed to them.

I also argue against a reduction of intentional states to the neurological mechanisms. My view is that the intentionality is institutional, and is conferred on actions that count as moves within norm-governed practices. It is only derivatively on internal neurological states. This, like the intentionality that we attribute to linguistic expressions, is abstracted from the kinds of actions (including linguistic ones) they enable the agent possessing them to perform, and derived from the norm-governed practices in which such actions have their life.

I conclude with a naturalistic account of the norms and practices within which human actions count as having content (reasons). I do not give a naturalistic *justification* for these norms, but a naturalistic *explanation* for how normativity in general arises (whatever the norms happen to be). I appeal to

forces of natural selection operating on groups, and (often tacit) practices as shared and enforced ways of acting that enabled different practices and the groups that practice them, to survive and prosper. This feeds into the crucial step in explaining intentionality naturalistically: explaining how the ability to attribute intentional states to others evolved alongside, and made possible, human linguistic interactions. It also evolved alongside the ability to attribute to oneself the intentional states that others are licensed to attribute to you, and the disposition to live up to this "self-conception" as someone with those beliefs, goals, and desires.

Acknowledgements

Thanks are due to my supervisors Jeff Pelletier and Wes Cooper for feedback, support and encouragement, in spite of their disagreement with much of what I have to say. Thanks also to Thomas Wheaton Bestor, who instilled in me the perspective from which I speak. Thanks to my examining committee, Mike Dawson, Cresssida Heyes, Don Kiuken, and Jeff Foss, for insightful questions and acute editorial comments. Thanks also to Tal Aviran, Lisa Kline, Patrick McGivern, Caireen Ryan, Krzysztof Swiatek, and many other friends and relations for listening to my ramblings, reading early drafts of chapters and providing helpful feedback. Thanks to Lark Clark and Cathy Ennis for providing the soundtrack to my early investigations. Lastly, but certainly not least, thanks especially to my wife-to-be, Kristine, for endless patience, tireless proofreading, and unfailing support.

Table of Contents

List of Figures

Intentionality in Action: Looking for "Life" in All the Wrong Places

Frege's idea could be expressed thus: the propositions of mathematics, if they were just complexes of dashes, would be utterly dead and utterly uninteresting, whereas they have a kind of life. And the same could be said of any proposition: Without a sense, or without the thought, a proposition would be an utterly dead and trivial thing. And further it seems clear that no adding of inorganic signs can make the proposition live. And the conclusion one draws from this is that what must be added to dead signs in order to make a live proposition is something immaterial, with properties different from all mere signs.

But if we had to name something which is the life of a sign, we should have to say that it was its use.

—Ludwig Wittgenstein (1958/1933–6)

I The Cartesian Dilemma

Nowadays it's difficult to find any people who would call themselves Cartesian dualists. This idea that a human being is a concatenation of a physical body with a non-physical mind whose operation is beyond the Laws of Nature and which cannot be studied scientifically, is no longer appealing to most people; especially with the growing faith in the Progress of Science over the last century or so. It seems to most who consider this, that the alternative to Cartesian Dualism must be some form of contemporary physicalism, where the mind is instead viewed as a physical system whose operations can be studied scientifically.

I want to make two points about this dilemma. The minor point is that *both* horns are uncomfortably sharp. Dualist accounts, positing the mind as a non-physical, non-spatial entity, are obviously out of favour in today's scientific world-view. But in rejecting Cartesian dualism, physicalism has kept a lot of the Cartesian framework, in particular the fundamental assumption that my essence is as a thinking thing (this time a brain) attached to and controlling a non-thinking thing (my body). By keeping this framework, physicalist approaches to the mind, and "information processing" approaches to cognitive science, simply exchange a mind-body dualism for a brain-body dualism, and so inherit many problems that plague dualist accounts. In addition, they face new problems about which Cartesians were never that concerned; about how a physical thing (like a brain-state) can have mental properties (like intentionality or consciousness). It's these problems that make the physicalist horn of the dilemma as uncomfortably sharp as the Cartesian dualist horn.

The major point I want to make about this dilemma is this: it appears that with Cartesian dualism falling from favour, physicalism of some form is the only (respectable) game in town, but this is not so. It's possible to avoid both horns of the above dilemma, by rejecting a central assumption made by both dualist and physicalist approaches to the mind. Many philosophers of mind and language, linguists, psychologists and cognitive scientists are indeed playing a rather respectable different game.

This different game begins by rejecting the assumption that Descartes uses to get the dualist enterprise started, the assumption that physicalism unquestioningly inherits. Descartes begins asking, What can I know? and rightly concludes that I can be certain that I am thinking, that I am doubting, that I am having such and such sense-impressions. He then concludes that this confirms to me something else that I can know with certainty: I know that I exist. But then Descartes asks the metaphysical question: What am I, this thing which I know to exist? And Descartes deduces that my *essence* is as a thinking, non-bodily thing. This is because, to Descartes, "I am thinking" does entail that I exist, while "I have a body" doesn't entail that I exist. It doesn't entail this because of Descartes' assumption that it's reasonable to doubt the truth of "I have a body". Descartes believed that my thoughts could be the way they are even if this were false.

This assumption that the mind is a thinking thing that is at least conceptually separable from the body defines the whole dualist project, and, as I'll argue in the next chapter, it defines the physicalist project that followed it. Yet this assumption takes back the move that got the doubt about bodies started in the first place. Let me explain. Descartes begins by reminding himself that "...I must nevertheless here consider that I am a man, and consequently, I am in the habit of sleeping..." and in sleep, of dreaming. And he notes that these dreams occur when he is "...lying undressed in bed".[1] His dreaming argument is one of

[1] Descartes (1641), second page of *Meditation* I; p. 19 by Adam and Tannery pagination.

the devices he uses to support the above assumption, that "I have a body" could be false. Descartes argues that *I am walking* would not entail that I exist, unless it were true that I'm walking. And I cannot know that I am walking; this would depend on the movements of my body, which "sometimes does not exist, as in dreams, when nevertheless I appear to walk."[2]

However, if I'm wrong about what I think I'm doing, then even on Descartes' descriptions, this is because *I'm actually doing something else.* I could instead be asleep in bed and dreaming. For instance, right now I believe that I am sitting in front of my computer, writing. But it may be false that I'm sitting at my computer, writing. For all I know I might instead be lying in bed undressed, asleep and *dreaming* that I am sitting at my computer, writing. Goodness knows, I dream about this often enough lately. But with either of these alternatives, I am *doing* something: either I'm sitting at my desk, writing, or I'm lying in my bed, sleeping and dreaming. I might seriously doubt that I really am doing what I think I'm doing. I may very well not be doing that. But I am always doing something.

While I cannot validly infer from the premise "I think that I am writing" to the conclusion "I am writing", Descartes argues, "I think that I am writing" does entail that I, who am thinking this, exist. But what is the "I" that I think is writing? As John Cook (1969) points out,

> ...everything in Descartes' *Meditations* is said under the supposition that he may be dreaming. Whatever sort of philosophical doubt this may raise, there is at least one thing certain; if he should ask himself 'What am I?', he can answer that he is a man who sleeps, undressed and in bed, and often dreams. (p. 123)

The "I" that I think is writing is the entity that might instead *be lying undressed in bed asleep* and dreaming that I am writing.

The move from epistemology to metaphysics leads us seriously astray. Sometimes, Descartes correctly argued, I am wrong about what I think I am doing. But when I am wrong about what I think I am doing, it's because I —a whole embodied person— am doing something other than what I think I am doing. Thus, argues Cook (p. 122), Descartes begins to separate himself from his body and the world, by reminding the reader about his being a human being;

[2] Descartes(1911), Volume II, p. 207.

embodied, and acting in a world. "To take back this beginning," says Cook, "is to take back everything" (p. 123).

II My Essence is as an embodied Agent

The other game I referred to just now takes this as the starting point. Here, if we move from doing epistemology to doing metaphysics, and move from asking 'What can I be certain of?' to asking 'What am I?', then rather than concluding '*sum res cogitans*,' we might better conclude '*sum res agens*.' My essence is as an agent, a whole embodied person who does things. Rather than seeing a human being as principally a thinking thing, in control of and connected to a body, this perspective views a person as an essentially embodied, world-embedded, socially situated agent. (But an agent who can do things that many other creatures can't do; especially "mental" things like planing what to do next Thursday evening.) Because of this focus, I'll refer to such approaches under the rubric "embodied action" cognitive science. It's also known as situated action, enactive cognitive science, and ecological cognitive science; when I use "embodied action cognitive science" or "embodied approach to cognition", I implicitly invoke these other brands as well.

In this dissertation, I'll begin by explaining my claim that physicalism, being a reaction to Cartesian Dualism in which non-physical minds are replaced by physical brains but little else changes, inherits many of the problems that plague dualism. The central problem is that physicalism, like dualism, takes the central feature of human beings to be that we are essentially *thinking things*. I'll then sketch a picture of an embodied approach to the study of cognition that takes as the starting point the assumption that human beings are essentially agents acting and interacting with other agents within a social and physical world.

By suggesting alternative assumptions to guide our research, however, I'm not out to reject all of physicalist philosophy of mind, nor all of "information processing" cognitive science. I am out to question and to change some of the fundamental assumptions that guide research in these fields, though. These assumptions are not defended, because they are so much a background to the language-games in which the arguments within these fields have their life. Ludwig Wittgenstein gives us two metaphors to illustrate the way such assumptions function. One is the notion of a picture:

> 115. A *picture* held us captive. And we could not get outside of it, for it lay in our language and language seemed to repeat it to us inexorably (1958).

One way to look at what embodied cognitive science is doing is that it is giving us an alternative picture to frame our inquiries within.

Another way to look at these assumptions, and thus at what embodied action cognitive science is doing, is given by Wittgenstein's metaphor of the assumptions as part of a riverbed, through which the river of our inquiries flows, shaping and constraining the course of these inquiries:

> 96. It might be imagined that some propositions, of the form of empirical propositions, were hardened and functioned as channels for such empirical propositions as were not hardened but fluid; and that this relation altered with time, in that fluid propositions hardened, and hard ones became fluid (1969).

The point then, is to try to alter some of the hardened propositions, to turn them back to fluid and allow our inquiries to flow in slightly different directions. The metaphor of the picture that holds us captive recurs within this metaphor, when Wittgenstein talks about the mythology:

> 97. The mythology may change back into a state of flux, the riverbed of thoughts may shift. But I distinguish between the movement of the waters on the river-bed and the shift of the bed itself; though there is not a sharp division of the one from the other... (*Ibid*).

The waters of inquiry, then, may wash away bits of the riverbed with them. It appears that for inquiries in information processing cognitive science the waters may flow quickly. It's important to ask which parts of the solid riverbed they flow past, and whether this fast-paced inquiry makes changes to the parts of the riverbed that need to be moved. I believe, in fact, that inquiries in information processing cognitive science have maneuvered well away from a particularly troubling and apparently intractable boulder that needs to be shifted. It's a boulder that philosophy of mind has acknowledged impedes several promising looking paths that inquiry could flow down. Philosophers of mind have been pushing directly against this boulder for quite a while; it still seems rather stubborn and immovable.

Rather than joining the push directly against this rock, I'm going to try a different tactic, also recommended by Wittgenstein:

> Scraping away mortar is much easier than moving a stone. Well, you have to do one before you can do the other.[3]

Wittgenstein continues his riverbed metaphor with the following:

> 99. And the bank of the river consists partly of hard rock, subject to no alteration or only to an imperceptible one, partly of sand, which now in one place and now in another gets washed away or deposited (1969).

The way to begin to shift that boulder, then, is to work a little more indirectly, agitating away at the sandy parts of the bank that the rock is embedded within. In order to shift some of the rocks that impede progress, I aim to direct my inquiry into the sandy riverbank in the places surrounding that particularly intractable boulder.

The apparently intractable boulder is the problem of naturalizing intentionality. Physicalist philosophy of mind has been pushing at it for quite a while, but it's proved very difficulty to budge. Some aspects of embodied action cognitive science have done a great deal to agitate around the area (although some antirepresentationalists have simply argued that the boulder isn't there at all). I'm not going to give a thorough defense of embodied cognitive science here; that could take several large books.[4] Here I'm going to adopt some of their tools and methodologies, to actively concentrate my agitation in the sands that hold this problem in place. By doing so, I aim to show that this boulder isn't all that hard to move after all, if it is approached a little more indirectly.

III *An Embodied Action Approach to Intentionality*

According to physicalist philosophy of mind, intentionality is a special feature of certain special items. These items —items like words, pictures, and special "mental" states and/or processes called "representations"— are held to have the special property of meaning, being about, being directed at, or representing other items or states of affairs.

The scene is set by the extract from Wittgenstein's *Blue and Brown Books* that I used as the epigraph at the beginning of this introduction. What is it that gives a "dead" sign its "life"? he asks. A theory of intentionality attempts to

[3] This is from Wittgenstein's notebooks (a note dated 1940), published as *Culture and Value* (1980), p. 39.

[4] It has. In Section 1.3 I'll give many examples.

explain how such items like marks on paper, noises, and bits of neurological matter "come alive" like this, so that rather than being "dead" ink-marks or noises or brain-states, they have intentionality; they have these special properties of being about, being directed towards, representing, meaning, or referring to other items or states of affairs. The principal questions for a theory of intentionality are these: (1) What makes item A "come alive" to be the sort of thing that has intentionality? (2) What makes item A about item B (rather than some other item)? (3) How does the intentionality of a part of an item (like a sentence, or a representation) contribute to the intentionality of the whole?[5] Answering these questions from a physicalist approach has so far proved notoriously difficult to do. Nonetheless, many researchers —especially cognitive scientists— have faith that these questions about how items have intentionality and can represent will one day be answered. They proceed assuming that it's unproblematic to simply assume that there are representations in people's brains, that these representations have contents (that determine which things they correctly represent), and that cognition can be explained —some say entirely— in terms of processes operating on such representations. The question of *how* exactly these representations can be said to correctly represent certain things and to misrepresent others is left in the "too hard" basket for someone else to answer.

Considering the vast number of different —so far unconvincing— attempts to answer these questions,[6] I've begun to wonder if these are the best questions to ask. I've come to the conclusion that they are not. We've been looking for the "life" in the wrong places. This dissertation is a beginning attempt to show why I think this is so, and to present some alternative questions. I'll show the problems of intentionality to be an artifact of the representational theory of mind, in which there are mental items that exist in minds (neurological items or states that exist in brains for physicalists) and in which such items have the property of being about worldly items. If we are

[5] Cummins (1989) calls the first two the problem of representations, and the problem of representation, respectively. That is, the problem of representations is: what makes certain items represent? The problem of representation is: what make an item that represents, represent one (sort of) item and not others? The third is often referred to as the problem of compositionality.

[6] My own answers to these questions in my Master's Thesis was one of the factors that began to convince me of the futility of attempting to answer these questions, and one of the factors that initiated my suspicion that the questions themselves incorporated misleading assumptions.

going to find "life" anywhere, I will argue, it is not to be found in *items*, but in people's *actions*.[7] People's actions count as being directed at objects and states of affairs, by virtue of the norms of the practices these actions are moves within.

IV Where are we going, and how are we going to get there?

In Chapter One, I outline some of the problems that traditional "information processing" approaches to cognitive science and physicalist philosophy of mind inherit from the Cartesian tradition. I then briefly present a general outline of the central tenets of embodied action approaches to cognition.

In Chapter Two, I'll spend some time discussing the role of representations both in traditional and in embodied cognitive science, and the challenge from some embodied action camps that cognition doesn't involve representations at all. I'll then illustrate the central problem for representational cognitive science: the problem of giving a naturalistic account of the *content* of representations. Such an explanation is necessary to account for the undeniable fact that a representational system can get things wrong; we human beings all too often represent things to be other than as they are (e.g. I take a skunk-on-a-dark-night to be a cat). My point here is that the only way to account for misrepresentation is to accept the fact that content is a normative property. The representation *should* represent certain things and not others. Thus the problem of naturalizing content is not to be *solved* by reducing content to physical properties. Justifying an "ought" in terms of an "is" cannot be done. Rather the problem is to be *avoided* by embracing the normative nature of the content of representations. The task is to explain naturalistically, not particular norms and thus particular representational contents, but how norms and intentionality arise. I will tell a *Just So Story* about how the ability to follow norms (whatever those norms happen to be) and the ability to attribute intentionality evolved. Chapters Three, Four and Five embrace this normative nature of intentionality. Chapter Six does the explaining.

In Chapter Three, I present different views on the relationship between the intentionality of language and the intentionality of people's mental states. I argue that a background of tacit norms and practices is a precondition for intentionality, and that most animals operate purely at this background level.

[7] An early version of this thesis that intentionality attaches to actions not to items is from Bestor's (1990) interpretation of J. L. Austin, as propounding the view that semantic properties attach to speech acts, not to words.

Human practices and norm-following behaviour arise out of such pre-conscious background capacities. I also argue that all *contentful* intentionality (not just object-directed, but directed at the object *as a particular kind of object*) relies on language to make the contents explicit.

To explain the evolution of intentionality, then, I'll need to explain the evolution of language. The two go together. If this is so, then we need a good picture of what language is, and of the nature of the linguistic skills that I'm going to give this *Just So Story* about. In Chapter Four I outline and defend an embodied action (speech-act) approach to language and people's linguistic abilities. Here language is seen as a social practice, an activity that people participate in together. Thus the abilities that enable people to perform and interpret speech acts depend upon the ability to recognise a speaker's reason for doing what they did. And thus speakers must have the ability to use words in such a way as to make their reasons for doing that recognizable. The norms of the practice that the speech acts are moves within enable this to happen.

I argue in Chapter Five that rather than looking at the truth conditions on statements of folk psychology (statements that attribute beliefs, desires intentions etc. to people), we should look at the norms that structure the *practice* of attributing such intentional states to people. The justification for attributing an intentional state to someone, I will argue, does not depend on the presence of actual intentional states in the person (neither mental states nor neurological ones). Rather, the justification for attributing an intentional state to someone is the way the agent acts: how the agent has, does and will behave, talk, react, and so on. The attribution of intentional states to others is subject to the shared norms governing the practice of using public expressions for attributing intentional states to others. These publicly shared norms specify what must be the case (public events, actions, speech acts, etc.) for an attribution of an intentional state to someone to be felicitous. They specify the conditions whereby –within our shared practices and the norms that govern them– a person *counts as* being in a particular intentional state. The trick to this, the one that makes the whole system so darn useful, is that these norms work in reverse as well. In addition to ruling certain attributions of intentional states as appropriate, they also rule certain actions expectable (ceteris paribus) of agents to whom such states are attributed. Acting in ways that license others to attribute intentional states to you, also licenses them to expect certain behaviours from you. Because most of us accept the general injunction that one should act "rationally", acting in certain ways puts you under a social obligation to act in ways consistent with the intentional state people should attribute to you. You *should* act in the ways that someone with those intentional states should act.

The focus of all of this is on people's actions and the practices these actions are situated within, not on their internal states. For language, the focus is on speech acts and the public norms that speech acts meet when, within a particular social practice, they count as being directed at certain objects or states of affairs. Thus linguistic intentionality is a normatively instituted property of people's purposeful speech acts, not a property of the items that enable such actions to take place. Analogously, the neurological items that enable an agent to perform actions do not *themselves* represent. If anything counts as being directed at worldly items, it is the practice-situated actions an agent performs, not the items produced and employed in such actions; in spite of the fact that the actions could not take place without those items.

The task of Chapter Six, then, is to give a naturalistic account of the norms and practices within which human actions count as being directed at items and states of affairs and within which it is appropriate to attribute particular intentional states to people who behave in certain ways. This is not to give a naturalistic *justification* for these norms; Hume has shown the futility of trying to derive particular norms from facts about how things are. The aim is to give an evolutionary explanation for how norms in general arise (whatever those norms happen to be). This feeds into the crucial step in explaining intentionality naturalistically: explaining how the ability to attribute intentional states to others evolved. I will argue that it evolved alongside, and made possible, human languages and the kind of higher level intentionality evident in human linguistic interactions. It also evolves alongside the ability to attribute to oneself the intentional states that others are licensed to attribute to you, and the disposition to live up to this "self-conception" as someone with those beliefs, goals, and desires. Thus I end with an explanation of an important –and often missed—step in the explanation of how language evolved: an explanation of how the practice of attributing intentional states to others enabled the evolution of language.

Chapter One

An Embodied Action Approach to Cognition

When, O monk, the view prevails that the soul and body are identical, there is no salvation;
when, O monk, the view prevails that the soul is one and the body another, then also there is no salvation.

— Buddha (Sidhartha Gautama)

To confront the undivided mystery undivided, that is the primal condition of salvation.

—Martin Buber[8]

1.1 *The Cartesian inheritance*

This dissertation is directed at challenging, and offering an alternative to, some of the fundamental assumptions that unite many of the variant systems within contemporary philosophy of mind and information processing cognitive science. Alfred North Whitehead has some advice for people engaged in a task like this:

> When you are criticizing the philosophy of an epoch, do not chiefly direct your attention to those intellectual positions which its exponents feel it necessary to defend. There will be some fundamental assumptions which adherents of all the variant systems within the epoch unconsciously presuppose. Such assumptions appear so obvious that people do not know what they are assuming, because no other way of putting things has ever occurred to them. (Whitehead 1925, p. 71)

As Whitehead advises, I will not be directly attending to many of the intellectual positions that its proponents feel it necessary to defend. Many of the objectors in such disputes presuppose fundamental assumptions that, for me, are more important to attend to. The first part of this chapter is directed at illuminating some of these assumptions *shared* by both dualists and physicalists (and thus by

[8] This comment, and the preceding quote from the Buddha it refers to, are both taken from Buber (1923/1970, p. 138)

most cognitive scientists). I'll begin with a brief summary of these common assumptions and common problems in the conception of what it is to be a human being shared by the Cartesian and physicalist viewpoints. In the latter part of this chapter, I'll illustrate how the "embodied action" approach's conception departs from these assumptions, and the different set of "fundamental assumptions" that guide this collection of approaches.

Our story begins with Descartes, although there are certainly precursors of the separation of mind from body much earlier. Descartes bifurcated human beings into two parts: a mental thinking part, and a bodily acting part. The thinking part of a person, to Descartes, has quite different properties than the physical bodily part; the body has size and shape and location, is made of matter, and is subject to physical laws. The mind is non-physical, has no size, shape nor location and is not subject to physical laws.

Physicalists, in general, reject the thesis that the human mind is a non-physical object exempt from the laws of physics. While there are many different ways of being a "physicalist", making it difficult to cast all physicalists under the same description, there are commonalities between the major positions. In what follows I will outline some of the major Cartesian-influenced assumptions that physicalists accept.

1.1.1 The mind is nothing but the brain

Identity theorists, reductionists, eliminative materialists, and perhaps even functionalists all agree in accepting some interpretation of the statement that "the human mind is (nothing but) the brain." There is disagreement about how the "nothing but" should be unpacked, but this is disagreement within a large amount of agreement. For most "type" identity theorists, the statement translates to "the mind is just the brain, and there's nothing mental left unaccounted for". For most "token" identity theorists it translates to "Each token of mental process A is nothing over and above a token of brain process B". Other "token" identity theories, such as Davidson's anomalous monism, translate the "nothing but" into a claim that for each phenomenon referred to by a mental description, the very same process can be referred to by a physical description with no remainder (although for Davidson, neither description is reducible to the other). Reductionists (at least those who aren't identity theorists) unpack the above claim as "the operations of the mind can be explained by referring to brain processes and nothing but brain-processes". For

Eliminative materialists, it is unpacked to mean something like "'Minds' are a theoretical construct, that, like caloric fluid, science has proved don't exist. Everything we attributed to the operations of the mind can (or will be) explainable in terms of the operations of the brain." For functionalists mind is identified by what it does rather than by what it is. So for them, the statement turns into a claim of the form "The mind is whatever causes (intelligent) behavior. Science has found that the cause of (intelligent) behaviour in human beings is the human brain."

Thus all these different types of physicalists reject the notion of the mind being a *non-physical* thinking thing whose operation falls outside of the scope of physical laws. However, many still keep the Cartesian idea that the mind is a *thinking thing*, connected to a (Cartesian, non-thinking) body. The main change (some might say the only change) is that the thinking thing is now thought of as a physical brain (along with the central nervous system), rather than a non-physical mind.

1.1.2 My essence is as a thinking thing

On the Cartesian conception, to be a "thinking thing" is the "essence" of being human. My self is my mind. This thinking thing that I am happens to be attached to a physical body, but this physical body is not any part of my "essence". Because it is supposed that the body does not contribute anything —apart from inputs— to the operations of the mind, Descartes thought it sensible to doubt that I even have a body. The possibility that I am a disembodied mind was fuelled by the possibility of deception by the Evil Demon. It's possible, says Descartes, that the Evil Demon could deceive me into believing that I have a body, when I in fact do not.

In physicalist approaches, this conception of a person as "essentially" a thinking thing –a brain– has changed little. The body (that which is not the brain) is often not seen as an essential part of a person; the thinking part (brain) is the essential part, and the operations of the bodily part are not essential to the operations of the thinking part. The brain is physical, however, and depends physically on the operations of the body (e.g. for oxygenated blood), but these functions, and those of providing input and receiving output (theoretically, at least) could be replaced. For Cartesian dualists, even though I believe that I have a body, it's possible that I am in reality only a mind fed sensations by the Evil Demon. In physicalism, its similarly possible that I could be a disembodied "thinking thing" –a brain in a vat— fed "sensory" inputs by evil (or sometimes benevolent) neuroscientists.[9]

1.1.3 *This thinking thing interacts with a non-thinking body*

Thus while physicalists have rejected the idea of a Cartesian mind, they have preserved the idea of a Cartesian body (as a non-thinking thing, which is controlled by the thinking part of a person). This non-thinking part is controlled by a physical brain instead of a Cartesian mind, but little else has changed. The body is basically a vehicle for carrying the mind (brain, self) about, and for transducing information from the world, and for carrying out the thinking part's volitions.

The Cartesian picture of interaction between mind and body is evident in physicalists' treatments of both perception and action. Physicalists' position that the mind is in fact a physical brain removes the Cartesian problem of explaining the interaction between non-physical minds and physical bodies, but the interactionist picture remains. Here perception and action are usually viewed as separate, two-part processes (respectively, the input and output of the "information processing" operations of the brain). In perception, a mental representation results from sensory stimulation. In action, an event of mental willing (a brain-process) precedes or causes the physical, bodily action. A *purposeful* action is one that is caused by special kinds of mental (i.e. brain) processes and states, and a representation is a special mental (neurological) effect of stimulation of the sensory systems.

1.1.4 *Actions involve bodily movements*

Because of this "interactionist" conception, we get a dichotomy between acting and thinking. Thinking happens in minds. Acting is a mental process causing a bodily movement. An action, therefore, must involve a bodily movement. If it doesn't, then it's just a thought (and not an action). Thus "mental action" is almost a contradiction in terms. Anything that happens only in an agent's mind (such as deciding, planning, adding numbers) is not something the agent *does*. An action is a bodily *effect* of a mental volition. Mental phenomena (like "making

[9] See for example, Putnam (1981) and Zuboff (1996). I'm not denying the logical possibility of the brain in a vat scenario. (Although I do think that a person's interactions with other people entail that the inputs and outputs would have to be far more complicated than traditionally supposed; the "vat" would have to contain a world of virtual people –I don't think they could just be artificial intelligences– for the subject to interact with.) Here I simply want to point out the Cartesian picture, of my essence being my "thinking part", underlying this (alleged) possibility.

up my mind" about my next move in the chess game, or planning what to cook for dinner tonight) do not count as actions.

1.1.5 Mental phenomena are mental items

Partly because of this tendency to think of actions as bodily (not mental) phenomena, physicalists describe mental phenomena as states rather than activities. Rather than describing a person as *believing* something, they say that the person *has a belief* (in their mind/brain). Rather than a person expecting someone, they describe the person as having (their mind/brain containing) an expectation; their desiring something is described as their having (their mind/brain containing) a desire. This descriptive bias in favour of nouns rather than verbs —mental items rather than mental activities—shifts the emphasis from what I (a person) *do* to what I (a brain) *have.* And once this emphasis is shifted from what I do to what I have —to the fact that I do indeed have, for instance, a belief that it's going to rain— then it seems sensible to inquire about the nature of this belief that it's going to rain. The belief itself is assumed to be a thing, and since it can't be a thing in a mind, we ask "what kind of thing is it *really*? It's assumed that it must really be a state or structure or item in my brain somewhere, whose nature neuroscience can supposedly investigate and tell us about.

The ensuing view of mental entities in terms of mental *states*, rather than mental actions, abilities, dispositions and capacities, is perhaps also a consequence of taking the computer metaphor of the mind too seriously. Computers' operations are often described using state transition diagrams, diagrams that document the ways that inputs affect finite state machines, causing transitions from one state to another. Many approaches that take this metaphor seriously use similar descriptive devices to account for people's mental operations.[10] Thus physicalists often view mental processes as a sequence of states. People change mental states in response to "input" by jumping from being *in* one static mental state to being *in* another mental state. Having one propositional attitude together with a particular sensory "input", causes me to have another propositional attitude.

1.1.6 An artificial intelligence doesn't need a body, just inputs and outputs

This bifurcated view of a person as essentially a thinking thing, that happens to be connected to a non-thinking body, also drives much research in Artificial

[10] Examples include Fodor (1981), p. 120 and Putnam (1975b), p. 434.

Intelligence, where the aim is to create an artificial "thinking thing"; an artificial brain, or something functionally equivalent, or at least functionally similar, to one. The need for such an artificial intelligence to have a body (in addition to the physical apparatus that instantiates the thinking thing), is secondary. It's supposed that it's not essential to being intelligent that one have a body with which one can act in the world.[11] After all, to be intelligent one merely needs to be able to think. A linguistic interface with the world is assumed to be adequate for such a thinking thing, so that it can have information imparted to it and so it could "voice" its thoughts on the information so imparted.

1.1.7 Cognition is thinking, and thinking is information processing

Likewise in most approaches to cognitive science, the thinking part of a person is seen as centrally important; the bodily, non-thinking, part is peripheral. The sensory systems' are thought of as "input" to the brain, the "outputs" of which are "commands" to the motor-control areas of the brain (the actual limbs and muscles are even more peripheral). Cognition is what happens in between input and output. In cognitive science the prevalent assumption that guides all research is that all human cognition can be explained in terms of computational[12] "information processing" processes operating on represented information.[13] Sharing this assumption, says Michael Dawson (1998, p. 4-6), is the factor that unites all cognitive science researchers from the various disciplines that contribute to cognitive science. In spite of many disagreements within cognitive science research, he says, cognitive scientists can at least understand what one

[11] Compare Dreyfus' criticism of the CYC project in his introduction to his (1992) *What Computers Still Can't Do*. Dreyfus claims that our ability to imagine feeling and doing things is what enables us to organize and understand verbally represented knowledge and descriptions of situations. Computers, not having bodies, don't have this ability. Dreyfus also deftly counters the objection that computers don't need arms, legs, eyes, and so on, since people like Madeline, a woman who is blind, from birth and paralyzed (and so cannot imagine seeing and doing things), can still learn from books read to her. His objection is to suggest, among other things, that "a person's bodily skills and imagination are a necessary condition for acquiring common sense, even from books". See the introduction, p. xx for details.

[12] I use "computational" here in that sense that many connectionist network models of cognitive processes are also computational. They also take inputs, and do something "computational" to them to produce outputs.

[13] This claim is made by advocates of both classical (e.g. Fodor and Pylyshyn 1988) and connectionist (e.g. Smolensky 1988) approaches to cognition. See Von Eckardt (1993) for a defense of this as a methodological assumption within the representational and computational approach to cognitive science. For defense and criticisms of this position, see Vera and Simon (1993), and the critical responses by situated action theorists in the same special issue of the journal *Cognitive Science*.

another is saying when they disagree, because they all share the assumption that cognition is information processing. Barbara Von Eckardt (1993) argues similarly, that all cognitive scientists endorse the assumption that "the cognitive capacities consist, to a large extent, of a system of computational and representational (i.e. information processing) capacities" (p. 53). Thus the more important and fundamental questions for cognitive scientists to ask of any cognitive system, are "what information processing problem is it solving?" and "how is it solving that information processing problem?"

Information processing is something that can be done by a brain (or by something functionally equivalent, like a computer or a connectionist network), something that isn't moving or acting, only "thinking". Thus cognition is seen as processes operating on the information represented by the inputs, to produce the outputs. And this information processing happens independently of any connection with a body, beyond the body's being a source of inputs and a destination for outputs.

1.1.8 Cognitive science explanations of people's abilities, must be information processing explanations

Dawson claims that this assumption does not limit cognitive science. It doesn't constitute a limitation, he says, because "explanations of information processors require many different kinds of descriptions" (p. 7). It does provide constraints and "narrows cognitive science's focus considerably." (p. 6) Dawson believes that this narrowing of the focus constrains, in a productive way, the kinds of explanations we can legitimately give in cognitive science. For instance, Dawson says that ,

> ...if we overhear a sentence that says "the mind has property X", and we know that this property is not true of information processors, then we also know that this is not a meaningful sentence in cognitive science (p. 6).

Thus "the mind is a non-physical thing, exempt from the laws of physics" is not a meaningful sentence in cognitive science. Von Eckardt presents her case very similarly, arguing that the assumptions that cognitive capacities are information processing capacities "constrain what counts as a possible answer to each of the basic questions." She continues:

> Thus, in endorsing the substantive assumptions, cognitive scientists limit themselves to entertaining only answers to the basic questions that are, roughly speaking, formulated in information processing terms (p. 53).

Thus, for Von Eckardt, answers to questions about human cognitive capacities, and what enables human beings to be able to do such things, will be phrased in terms of represented information, and information-processing operations on that information.

These constraints might not be entirely productive, however. On Dawson's account, a statement like "Each mind is essentially part of an organism with certain behavioural and sensory dispositions and capacities, situated in a particular ecological niche (*umwelt*)", and "the mind's fundamental operations involve the control of socially-situated, real-time action in such a niche", would not be meaningful statements in cognitive science, since these statements are not true of information processors. Or rather, if they are true of information processors, it seems that "information processors" is a concept that stretches beyond the narrow category of devices (such as digital computers and connectionist networks) that perform computations on represented information, devices that information processing views of cognition use as metaphors.

Furthermore, as Dawson concedes, "cognitive science will only be able to provide explanations of those phenomena that will yield to a representational approach." The embodied action approach to cognition that I'll be presenting in the next section disagrees. While *some* cognitive processes might productively be cast as (representational and computational) information processing processes, it seems at least premature –and possibly quite misleading, or even outright false– to assume that *all* cognition can be productively cast this way. As I'll show in the next few chapters, many promising answers to questions about human cognitive capacities are not phrased in terms of information processing, nor in terms of represented information. It seems overly narrow-minded (excuse the pun) to declare that such research is not *cognitive science* research.

1.2 Dualist Problems in Physicalism.

John Searle devotes the first chapter of *The Rediscovery of the Mind* to a similar (and much lengthier) diagnosis of the "conceptual dualism" inherited by physicalist approaches to the mind. Searle rightly notes that by accepting the Cartesian vocabulary and categories of mental and physical, mind and body, physicalism "accepts the terms in which Descartes set the debate" (p. 54) terms in which "mental" means "non-physical" and "physical" means "non-mental" and thus physicalism is "really a form of dualism" (p. 26). Thus contemporary physicalism is a revision of dualism, but in no large way is it different. We've graduated from talking about minds to talking about brains, but conceptually little else has changed. While I don't agree with all of Searle's claims, and think that he unfairly (mis)characterizes some physicalist approaches, at least in his diagnosis of the overall physicalist framework there is much in his conclusions that I do agree with (even if sometimes for different reasons than those he gives). Although Searle and I share a similar negative view of the current state of play, in which an undesirable physicalism is supposed to be the only (respectable) game in town, we differ considerably in our views of the alternative game we could go and play instead. I'm not going to over-complicate an already complicated analysis by attempting a detailed comparison with Searle, however. I'll simply point out some of the problems that physicalism inherits from dualism, and then go on to describe this "other game" I want to play instead.

The major problems in physicalist philosophy of mind are due to this Cartesian conception of "mental" meaning "non-bodily", and of "bodily" meaning "non-mental". These translate into "neurological" meaning "non-bodily", and of "bodily" meaning "non-neurological". We accept a separation between the (mental) thinking part of a person and the (non-mental) body, and then incur problems in figuring out the relationship between, and especially the interaction between, these two parts of a person.

1.2.1 Consciousness and subjectivity

First, there's the problem of giving a physical account of subjectivity and consciousness. People like Nagel (1979) and Jackson (1982) argue that physicalist accounts which try to reduce or explain mental events, states and processes in terms of physical events, states and processes, will necessarily leave something important out: subjective elements of "mental" experience. They argue that what an experience "feels" like to the person experiencing it, the subjectively experienced content or the phenomenal properties —the qualia— of those mental events, states and processes are essentially first-person phenomena. Thus they would necessarily be absent from physicalists' third person, scientific account of a human being's neurological processes.

The problem of explaining consciousness in terms of purely physical (e.g. neurological) processes, states and events is very similar to these concerns about subjectivity. It seems that if a description of mental processes in third person scientific neurological terms would leave out subjectivity, then it would also leave out consciousness.[14] We have little idea of what consciousness is, exactly. And we seem to have even less idea of how it might be given a physical (i.e non-mental), scientific, explanation. But we supposedly know that entities with minds (such as humans) have it, and things without minds (such as toaster ovens) do not have it.

1.2.2 *The problem of other minds*

Similarly, the problem of other minds is still a problem in the physicalists' framework, since the conscious part of a person (the brain) is seen as hidden "inside" in a realm inaccessible to other people (without surgery or using big fancy scanning devices). The (non-thinking) body is the only publicly observable part of a person. For instance, we have the problem of determining whether a non-human might be said to be "intelligent" or conscious. The problem is that something could behave as though it is conscious, yet it could still seem sensible to argue that it may not in fact have a mind or be conscious. The difficulty (perhaps impossibility) of proving whether or not something is genuinely conscious, complicates the picture even more. Imagine an artificial being, like Star Trek TNG's Commander Data, with a body like a human's, that behaved in many ways like a human would –that is, it behaves like it is genuinely conscious (thinks, has a mind). Of such a being, some (e.g. Searle) still believe that we can sensibly wonder: "Could an artificial being, who demonstrates all the abilities that usually go along with consciousness, truly be conscious, and have a mind?" Commander Data has a positronic brain, but does he have a conscious mind? Is it all just clever programming, which merely makes Data *behave as though* he has a mind, when really he is not "conscious" at all?

Should we at least wonder whether this is even a sensible position to hold? Is *having* a mind (a place where conscious thoughts happen) really the salient difference between Data and ourselves? If Data was a biological alien being, rather than an android, would such questions arise?

It's objections like this that move many people towards functionalism. For many functionalists there would be no difference between behavior that is

[14] See for example Elitzur (1989) and Chalmers (1996).

genuinely intelligent and something behaving *as if* it is intelligent. Functional equivalence is all the equivalence that matters. But many objections to functionalism are made on just these lines: that something could be *functionally* equivalent to a human being, but be a "Zombie", and not have mental states with subjective "phenomenal" properties, or qualia. [15] That is, behaving like a human being isn't all there is to being conscious. To be like a human being in the *relevant* respects, your internal states must also possess the phenomenal properties, the qualia, that human's internal states possess. (Dennett (1995b) takes the notion of a zombie to be self-refuting, calling it "a strangely attractive notion that sums up, in one leaden lump, almost everything that I think is wrong with current thinking about consciousness".)

John Searle's (e.g. 1990) Chinese Room argument is based on similar concerns. It is mostly an argument against the Turing Test, which holds that a good test of whether an artificial intelligence is truly "intelligent" is the ability to reliably, consistently, over a long period of time, converse in a way indistinguishable from the conversation of a human being. Searle argues that even if it were possible to program a digital computer such that the computer would pass the test, the computer would still not be *genuinely* thinking. Searle argues that something that behaves —that is, converses— as though it is intelligent could still lack that special "something" that human beings have and toaster ovens do not. This special something is intentionality (the true mark of "mental" operations, at least since Brentano 1874/1973). To have genuine intentionality is to have thoughts that are genuinely *about* things. These thoughts are mental operations that, on Searle's account, are emergent properties caused by neurological operations in brains. Thus behaviour is, to Searle, completely independent of the thoughts. Either could exist without the other.[16] To Searle there is this crucial difference between something genuinely having intentionality, and something behaving as though it has intentionality,

[15] Examples include Kirk (1974), and more recently, Chalmers (1996). Moody (1994) objects that there would be behavioural differences between zombies and conscious people, in their conversations about consciousness.

[16] Turing's test, on the other hand is based on the supposition that these are not separable. In order to reliably, consistently, over a long period of time, converse in a way indistinguishable from the conversation of a human being an entity would have to have a mind. Some interpret this test in a Rylean fashion (see (Ryle 1949), esp Chapter 2), that behaving in this way just is having a mind. When we say that something or someone "has a mind" we do not refer to having a special cause of its behaviour, rather, we say that it has the ability to do the kinds of things we say that only "en-minded" beings can do.

when in fact it does not. What makes it true that an entity has *genuine* intentionality, as opposed to simply behaving *as if* it has intentionality? The truth-condition of attributions of genuine intentionality is the presence of genuine thoughts in a mind causing the behaviour that licenses the attribution. To Searle, the behaviour must be *caused* by genuinely mental (i.e. high-level neurological) operations, rather than purely physical (i.e. non-mental, because non-neurological) operations such as those in toaster ovens and digital computers.

1.2.3 Problem of Cartesian bodies

Another symptom of the physicalists' positions is that they have for the most part inherited and unquestioningly accepted the idea of a Cartesian body: a *purely* physical (i.e. non-mental) acting part of a human being controlled by the thinking part. But a human body, so construed, is a fiction. No such thing exists. John Cook (1969), Antonio Damasio (1994), Frank Ebersole (1967), Douglas C. Long (1964), and Thomas Wheaton Bestor (1976) all argue the incoherence of conceptually separating human beings into dualistic entities, one of which is a human body, a purely physical (that is non-mental) entity, a component conceptually separable from the mind. Damasio, for instance, argues that the way we think is in many ways dependent on bodily functions. This is saying more than just that the operations of the mind depend on the operations of the brain. The body, he says, "contributes more than life support and modulatory effects to the brain. It contributes a *content* that is part and parcel of the workings of the normal mind" (226). Damasio argues that the way we think depends in many ways on "bodily" functions –chemical secretions of glands, for instance– that do not originate in the brain or nervous system. Furthermore, the body is a "ground" of all our representations of our environment; such representations are "engendered in the brain, on the basis of the body's anatomy and patterns of movement in the environment." (p. 235) To Damasio, the picture of a brain alone, or the body alone interacting with the world is a fiction. Mind and mental functions arise out of the whole entire organism, rather than out of a disembodied brain. (p. 229)

Bestor (1976) presents an argument for the converse side. Just as Damasio argues that cognitive functions are not purely in the brain, Bestor argues that bodily functions are not entirely non-mental. Bestor concludes with this:

> The dualist's conception of some purely bodily component of a human being — of a component identifiable by terms which make no direct reference to anything mental, which neither presuppose nor imply knowledge of the appropriateness of mental terms, and which leave epistemologically open the existence of other minds — is simply a figment of philosophical imagination. There is no such conceptually separable bodily component of a human being. There never has been. We aren't bodies plus minds. We are persons, agents, human beings. Hence, much as many philosophers have come to realise that what we normally speak of as mental states and processes are never really purely mental states and processes, so too we should now realise that what we normally speak of as bodily states and processes are never really purely bodily states and processes...
>
> The moral is obvious, but still, perhaps, a bitter pill to swallow. Philosophers blithely talk of human bodies and human bodily movements in all manner of contexts, rarely suspecting anything the least puzzling or problematic about such talk. They shouldn't. Save in a few special contexts, such talk is dualist talk. And, since the dualist conception of a bodily component is empty, such talk is usually empty too. (p. 24)

Frank Ebersole (1967) argues against the idea that an action partly consists of a bodily movement. The idea that bodily movements are simple things, and that actions are bodily movements plus some other stuff (mental volitions, scene-setting, rule-following, background capacities, and so on), says Ebersole, is a symptom of a peculiar set of philosophical presuppositions. "Bodily movements" are difficult to identify (Ebersole spends a large part of his paper failing to find an example of one), unless one has already accepted the Cartesian picture of a human being as a mind plus a (non-mental) body. "Of course," he says (p. 299), "nothing is more familiar or easy to talk about than actions. One must have a highly refined interest, and a highly technical vocabulary to talk of bodily movements." A similar phenomenon arises with the idea of pointing to a human body, compared with pointing to a person.

> Talk about bodies requires very special contexts. One has to know something about physical culture, art, girl-watching, undertaking, anatomical study, or police investigation before he knows how to point to a human body.... A person is not a body seen from a special point of view. Rather, a body is a person seen from a special point of view (Ebersole 1967, p. 303).

Many forms of physicalism keep this Cartesian concept of a human body as a fundamental entity in their ontologies. Bestor, Long and company argue that this is a mistake. It's a mistake because there are no *purely* bodily (that is, non-mental) states, processes, movements, or items. Just as there are no minds in the traditionally accepted Cartesian sense, there are also no bodies in the traditionally accepted Cartesian sense. Rather there are whole persons. And while person's have bodies, it's only by accepting the Cartesian categories that "body" can be used to refer to the non-mental, *purely* physical, part of a human being; that part that is not the brain and central nervous system.

1.2.4 The problem of naturalizing intentionality

The problem of giving a physicalist account of intentionality appears particularly intractable. For dualists, the mind contains thoughts. And one distinctive feature of mental phenomena like thoughts is that they are usually thoughts *about* things. As Brentano (1874/1973) famously points out, intentionality is a distinctive feature of "mental" phenomena. As a relation between a Cartesian mental item (such as an abstract general idea, in Locke's sense) and a physical object (or type of physical object) this relation of being *about* something is at least no more problematic than the interaction between non-physical Cartesian minds and physical bodies.

However, if mental phenomena are really nothing but physical phenomena, happening in brains not in Cartesian minds, then this property of *aboutness* that mental phenomena possess must also be explicable in physical terms. This requirement, that intentionality must be explicable in physical (i.e. non-mental) terms, gives rise to the problem of "naturalizing" intentionality. This is the problem of explaining in purely physical terms how a certain physical object (such as a brain-state) can stand to another physical object in this mysterious relation of being *about* that object. It's a problem especially since intentionality or "aboutness" is the property that, for Brentano, *distinguishes* mental phenomena from ordinary physical phenomena.

I'll concentrate on this problem of intentionality in this dissertation. I'm going to argue that it is a symptom of two moves, each of which I've mentioned above.[17] One is the separation of mind and body as distinct realms, and the problems incurred when physicalism insists on a physical (that is, non-mental)

[17] It is also a symptom of moves that I have not yet mentioned. The prominent one is the focus on the relation between individual agents (or individual minds, I should say) and the world.

account of the mental realm as Descartes conceived of it. The other is the reification of mental phenomena, from people's mental acts and dispositions into mental *items* that exist in a person's mind; the move from speaking of Fred wishing that it would not rain, to speaking of Fred having a wish that it would not rain, to speaking of Fred having a brain state that *is* his wish that it would not rain. These combine to constitute the problem that I will concentrate on.

This combination gives rise to Fodor's (1987) statement of the problem:

> Here are the ground rules. I want a naturalized theory of meaning; a theory that articulates in non-semantic and non-intentional terms, sufficient conditions for one bit of the world to be about (to express, represent, or be true of) another bit (p. 98).

The assumption that all researchers on naturalized intentionality take for granted is that the property of being about a bit of the world attaches to *physical* bits of the world. The problem of explaining how this can be, while the explanation uses no semantic or intentional terms (i.e. terms that themselves appeal to aboutness) has proved very difficult. The vast number of different attempts to explain this, and the objections to such accounts (usually showing that they do covertly make use of intentional concepts), all suggest to me that the problem –as it's currently phrased– is intractable. (I'll look at such arguments in more detail in the next chapter.)

I'm going to argue that this assumption that intentionality attaches to bits of the world is what makes the problem so intractable. I'll present an alternative view of the kinds of entities that possess intentionality, and show how it avoids the above problem, and provides a naturalizable account of intentionality. However, the account, since it does not explain how one *bit of the world* possesses intentionality, avoids meeting the problem head on, as Fodor's "ground rules" stipulate. It instead attempts to undermine the problem, so conceived, and thus to show the bankruptcy of attempts to solve it. This alternative account arises out of a recently emerging alternative to "information-processing" approaches to cognitive science, which is slowly gaining prominence.

1.3 Embodied Cognitive Science

As I mentioned earlier, most of contemporary cognitive science also works within this physicalist framework, in which all the interesting cognitive processes happen in the brain, and so a study of the brain *alone* and how it processes information will give us a complete picture of human cognition and how it operates. Michael Dawson (1998), as I said, maintains that cognitive scientists share at least the assumption that cognition is information processing. Comments like "The human mind is a complex system that receives, stores, retrieves, transforms, and transmits information,"[18] he says, are not uncommon in cognitive science texts. According to Dawson, agreement on this shared assumption unites cognitive scientists from varied disciplines and points of view; it enables cognitive scientists who disagree about the nature of the information processing to share information with each other, to speak a "common language".

For this approach, there are two main elements to cognition: representations of an external (objective) world, and computational processes operating on those representations. This is partially a consequence of cognitive science's ancestry, in both trying to understand the "mind" yet equating the brain with the mind. It's partially also a consequence of its ancestry in its constituent disciplines.

The metaphor of mind as computer has influenced cognitive science from the very beginning. And spurred on by early empirical success, this approach gained much favour. As Dreyfus notes with just a little cynicism, the program has had so much empirical success, that Newell and Simon's (1958, p. 6) prediction that "within ten years most theories in psychology will take the form of computer programs" has been partially fulfilled.[19] But as Searle (1994a, p. 855) notes, "'empirical success' is not enough to overcome conceptual confusion."

I'm reminded here of a joke which goes something like this:

> A man is walking home late one night, and notices someone obviously looking around for something on the pavement beneath a streetlight; he's picking up bits of trash and fallen leaves and peering underneath, he's poking around in the bushes at the edge of the path, looking fairly intently for something. The first guy asks "what are you looking for?"
>
> "I've lost my car keys," comes the reply.

[18] Dawson cites this as coming from Stillings et al (1987), p. 1.

[19] (Dreyfus 1972), p. 223, note 49. Dreyfus' Chapter 4 is a critique of this "Psychological Assumption" as he calls it, the assumption that the mind works (at some level of description) like a digital computer, and can be analyzed in terms of "information processing".

> So the first guy starts helping him look; peering around in the general area the other is searching. "Where do you think you dropped them?" he asks, after searching fruitlessly for a few minutes.
>
> "Oh, over there somewhere," the searcher replies, pointing along the street into the darkness beyond the pool of light cast by the streetlight.
>
> "So why look here?" the man asks, incredulous.
>
> "Because the light's better."

It seems that empirical success in the searching — we're able to methodically *look* in brains for cognition and cognitive processes (because the light's better) — has stimulated a lot of further research, with some impressive results. We've certainly learned a lot about brains and brain processes because of such research. But just because we're getting results in understanding the brain and brain processes, does this mean we're any closer to understanding human cognition in all its manifestations; to incorporating beliefs, emotions, intentionality, consciousness into our cognitive theories? These seem to be left as the "harder questions" to be explained once we learn more about other brain states and processes. Here I pause to give a reminder of Wittgenstein's (1958) commentary on this sort of outlook:

> We talk of processes and states and leave their nature undecided. Sometime perhaps we shall know more about them—we think. But that is just what commits us to a particular way of looking at the matter. For we have a definite idea of what it means to learn to know a process better. (§308)

Contemporary cognitive science has become committed to a particular (Cartesian) way of looking at cognition. The difficult questions, about intentionality, consciousness, subjectivity, are left as something we'll be able to explain once we understand the information-processing processes better. This seems, to me, to be an approach that encourages looking in the light, and downplays the importance of investigating thoroughly the context of cognition —or perhaps better, it's an approach that believes that a characterization of the context is unproblematic. I'm becoming more and more convinced that this is not the case. Our characterization of the context —what's "out there"— guides what we look for "inside". John T. Sanders (1996, §§31-32) puts it like this:

> ...while there are certainly fascinating things to be learned by investigating structure and internal mechanism, this route is not as vital, at least in the present

> state of the discipline, as is understanding ways in which we —not something inside us— behave, learn and act... There can be no doubt about the vital role played by the brain as we go about learning, searching, and acting. But the unit of analysis should be the organism, not the brain.... It is not brains that think, it is people. Things that go on in brains are necessary for cognition, but cognition could not go on in the absence of interaction with an environment, and environments could not be negotiated in the absence of bodies.

We may be more successful at explaining human "mental" processes then, if we approach cognition in a way that encourages a focus on characterizing the "outside" as well as the "inside" where we've been having so much empirical success. Recently some approaches to cognitive science have emerged whose scope encompasses a serious re-examination of our characterization of the context an agent acts in, by encouraging an approach that looks at the brain, body and world *together* as a system, rather than isolating the brain, and regarding the rest of the system as simply inputs and outputs to the brain. They focus also on the darkness outside that pool of light, on the physical, social and normative context in which all these brain processes do what they do. Many of them also encourage the search in the brain where we've been having all this encouraging empirical success (but also encourage using techniques for searching the dark areas to guide the way we might look at what's in the light).

This framework is referred to variously as "embodied", "enactive" "situated" or "ecological" cognitive science.[20] I haven't the space here to do much more than point to such approaches, and to give a brief outline of the basic motives and moves that unite this approach. Smith and Jones' (1986) introductory philosophy of mind text, introduces this kind of approach well. (The following introduction is adapted from this work, pp. 71-80.) They begin by asking about the difference between a stone and a seed. On the surface both are very similar. The major difference between the two is in terms of their potentialities: the seed has the capacity to germinate, grow and utilize the

[20] There are, of course, minor differences between these approaches. These are mostly differences in emphasis, however, rather then disagreements. Some do, for instance distinguish situatedness from embodiment, in that embodied creatures (e.g robots) could utilize detailed models of their environment to plan their actions, and then rely on such model-based plans to guide their behaviour. Such creatures would be embodied, but not situated, argue Pfeifer and Scheier (1999, p. 72). Most theorists who describe themselves as "embodied action" theorists, do however, assume situatedness as an accompaniment of embodiment.

elements of its environment for its growth and eventual reproduction. The seed has the potential for *life*. Now, two questions can be asked about things that are alive. One is the *conceptual* question, "What is required if a thing is to count as being alive?" The other is a *scientific* question, "What is it about a thing's inner workings and constitution that causally explains its particular form of life?" Philosophers are mostly interested in the former question, argue Smith and Jones (p. 72). And answers to the former question will probably influence the kinds of directions we explore when investigating the latter. Nobody nowadays would suggest that the former conceptual question is to be given in terms of some "vital spirits" present in living things. (This, we can presume, would prompt scientific investigations of what such vital spirits "really" are.) The answer to the conceptual question, they suggest, is instead to be given in terms of the object's potentialities: living objects have the potentiality for nutrition and reproduction. The scientific question is then to be answered in terms of the internal mechanisms, environmental factors, and possibly evolutionary forces, which causally explain these potentialities in such objects.

Now consider a similar question, about the difference between animals and plants. Again, the conceptual question will be one about what counts as an animal, what it is to be an animal. The answer will be given, not in terms of the animal's possessing a particular kind of thing ("animal spirits", for instance), but in terms of a difference in the entity's *capacities*. As a crude illustration, we could say that animals have the capacity for locomotion and the capacity for perception. The scientific question, then, is "What causally explains the fact that some things have these capacities?" rather than "What are animal spirits really?". This question is answered in terms of the internal physical workings of the organisms, and how those physical workings interact with the creature's environment, with perhaps some evolutionary explanations about how those internal workings came to be selected for.

The next step, obviously, is to ask about differences between human beings and animals. We can take this in two stages: first we might want to distinguish between creatures whose capacities lend themselves well to explanations in terms of the creature's intentions, beliefs and desires. What counts as an animal-with-a-mental-life? The answer is not to be given in terms of whether the animal has a particular special kind thing (e.g. a "mind") attached to it. Rather, the answer will be given in terms of what the animal can do, whether it has the capacity for certain sorts of rather more complex interactions with its environment. Research on animal minds indeed studies just this kind of phenomena. Researchers ask what kinds of interactions with their environment are certain (types of) animals capable of? The scientific question is thus going to be, again, "what causally explains these animals' capacities?", rather than "what is a mind, really?" The answer to this question will be given in terms of physiological, environmental, evolutionary and perhaps social mechanisms underlying these capacities (rather than answers about what a mind is, really).

The second step in explaining the difference between humans and animals is again going to ask questions about people's capacities, rather than about extra entities (*human* minds) they possess. The answer to the conceptual question of what counts as a person, will not refer to human minds, and determine what things have them. Rather, the answer will refer to the especially complex mental capacities that only human beings appear to have; for instance the ability to use language, our awareness of other's mental lives, our abilities to create complex tools, and the ability to seriously adapt our physical environment to better suit our perceived "needs". Something counts as a person if it has these kinds of capacities.[21] And thus the scientific question will not be "What is the human mind really?" but rather "What physiological, physical, social, environmental, and evolutionary mechanisms causally explain people's capacities to do the kinds of things we can do?"

It's this change in focus that unites the non-Cartesian approach that I'm working within. The change is from concern with minds, what they "really" are, and how the mind performs the tasks it performs, to concern with *creatures and their capacities*. We are especially concerned with figuring out what makes it possible for human beings to do the things they are capable of doing. And one important feature of this change in focus, is a move from a concern exclusively with the human brain and how it thinks. Rather, when explaining human beings and what they can do, we look at more than just brains and their physiology. We also look at the physiology of the whole human being, including but not limited to the brain. We also consider the role of social, environmental, and

[21] Of course, this is a crude answer, one that requires lots of caveats with respect to things we would like to count as people, but which fail in some of these capacities. Autistic people, brain-damaged people, people in a coma, and infants could all fail to count as persons according to such criteria. Of course, whether this distinction is important, will depend on the particular case, and for what purpose we wish to distinguish people from non-people; e.g. with regard to whether the entity's wishes or interests need to be considered, with regard to deciding whether it should be allowed to vote, etc.

evolutionary factors in explaining how it is that human beings are the kind of creatures that they are, and how such creatures are capable of doing the things they can do.

This overall approach, and its relation to the information-processing view of cognitive science is explained well by John T. Sanders (1996, §11-14) (well enough that it's worth quoting extensively):

> ...there are various ways to shed light on particular areas of inquiry, some of which stress trying to understand how things are constructed (i.e., what their parts are) and how they work taken one by one (i.e., how they work internally), and some of which stress how (perhaps) these very same things may be understood in their interactions in larger systems.
>
> [12] These are in no way inconsistent, it might be surprising to notice. They may very well call for precisely the same skills. It's just that, for any one topic (cognition, for example), what I am now going to refer to as the "analytic" approach (meaning not just "careful" but rather "focussing on parts and internal workings") suggests we look in one direction while the "ecological" approach suggests we look in another....
>
> [13] Where "analytic" and "ecological" approaches may actually bang into one another is on the issues of importance and value. Which approach is the *right* approach to some (given) area of inquiry? I do not believe that this is a question that has no answer. But I believe that the answer must be relativized to particular problem settings in a normal, non-controversial way. In short, whether an approach is the right approach depends upon your objectives. These different approaches plainly accomplish different ends, and thus must be evaluated in terms of the ends they aim at.
>
> [14] So in recommending an ecological approach to cognitive science, the claim (at least in my case) is just this: at the present time, under the circumstances of the problem situations that dominate the discipline (at least to the extent that there is any coherent direction that the discipline is taking), it is relatively more important to try to understand cognition in terms of its role in its broader environment than it is to try to further understand its internal construction and its basis in matter-energy. Indeed, as many have argued in the spirit of the mode of "analysis" (this time: "careful work") of the theory of natural selection, it may be that the details of the ecological picture will themselves provide clarity on many of the structural questions.

Rather than concentrating on what a human being is made up of, ("internal parts and their workings"), the unifying theme of this way of asking the scientific question is that the answers to the important questions about human capacities need to consider the human being as an inextricably embedded part of a larger system. The answers, then, will focus not on individual human beings and their internal parts, but on the larger systems that such individuals are part of, and will explain the way these overall systems functions.

Some attention will need to be paid to brain mechanisms, of course. Understanding cognitive mechanisms in the brain is a research project worth pursuing. Certainly, the brain plays a prominent role in most of an organism's activities. But —importantly— it's not everything. And it might be presumptuous to assume that all brain processes are all "information processing" processes. More important is to understand the broader context —of agents (inter)acting within their environments— within which these neurological processes play a role. So when we are looking at specific brain processes, it should always with an eye to the role they play in the overall physical and social interactive system of human lives, in which humans' cognitive capacities are displayed. As Sanders indicates, understanding this broader context of interaction may provide some guidance in our investigations of the brain-processes that facilitate such interaction.

Elements of the embodied cognitive science framework have been advanced by people in many different disciplines, approaching cognition from what is often referred to as a "non-Cartesian" perspective. The philosophical underpinnings come from people such as Ryle (1949), Wittgenstein (1958), Merleau-Ponty (1942, 1945), Heidegger (1927/1962), and more recently articulated by philosophers including Button *et al.* (1995), Dreyfus (1991, 1992, 1996), Dennett (1995a, 1996, 1991b, 1994), Haugeland (1995), Lakoff and Johnson (1980, 1999), Ó Nualláin (1995), Sanders (1996), Shannon (1993), Sprague (1999), Winograd and Flores (1986), and Wrathall and Kelley (1996). This "non-Cartesian" perspective is also at the core of autopoietic theories, founded by Maturana and Varela (Maturana, Mpodozis and Letelier 1995, Maturana 1975, Maturana and Varela 1980, Varela 1979), of which Varela, Thompson and Rosch (1991) is an offshoot, but by avoiding some of the rather technical vocabulary of autopoietic theories has perhaps been more influential, and is more accessible to newcomers to this approach. (See also Randall Whittaker's (1996a, 1996b) web-based introductions to autopoiesis and Maturana and Varela's work.)

Neuroscientists such as Damasio (1994) also articulate elements of this perspective. Work in "situated cognition" also advances and defends this way of thinking about cognition; for instance, see Clancey (1997), the papers in Kirshner and Whitson (1997), and the replies in the special issue of the journal *Cognitive Science* following Vera and Simon's (1993) critique of the idea that situated cognition poses a challenge to rules and representations style cognitive science. Ecological psychology, stemming from J. J. Gibson (1979) provides a theoretical base that many in this area draw upon. Embodied cognitive science also draws on developments in autonomous agent building, for instance the artificial agents designed and built by Rodney Brooks and his team at the MIT's AI lab (Brooks and Flynn 1989, 1991a, 1991b, Brooks and Stein 1993). It also draws on theoretical perspectives such as that voiced in Braitenberg's (1984) *Vehicles*. Pfeifer and Scheier's (1999) *Understanding Intelligence* is another good (textbook style) systematic introduction to the field of embodied cognitive science, with a focus on building autonomous agents, and the work of Brooks and Braitenberg. Pfeifer and Scheier also highlight the importance for this approach of recent work in artificial life, such as the articles in Brooks (1994) and Langton (1995). Approaches that focus on cognitive systems as complex dynamic systems, using insights from dynamic systems theory, also fit within this way of looking at cognition (see, for example, many of the contributors to Port (1995b) and also Abraham and Shaw (1992)). Developmental psychologists Thelen and Smith (1994) also draw on this dynamic system theory approach. A very comprehensive philosophical overview of embodied cognitive science, one that integrates many of these approaches, is Andy Clark's (1998) *Being There: Putting Brain, Body and World Together Again*. Clark's illustrative examples and explanations are part of what has brought this perspective together for me.

1.4 Methodological and Theoretical Assumptions of Embodied Cognition

Because of the focus on the overall system and how it operates, this "ecological" approach rejects the assumption that cognition is best thought of as information processing. (Remember that theorists such as Dawson (1998) and Von Eckardt (1993) claim that it's sharing this very assumption which unites the entire cognitive science discipline.) For most cognitive scientists, this is one of the (Lakatos-style) "hard core" (i.e. "don't question this") assumptions of the research programme that representational/ computational cognitive science is part of. Thus Pfeifer and Scheier's preface (1999, p. xiii) cautions readers that because of the insights offered by this new approach, their students reported that they "could no longer believe the kinds of explanations offered" in their classes in more traditional fields, particularly cognitive psychology. They add to the reader, "you have to be well aware that you may never be able to think about humans, animals, computers and robots in the same comfortable way as before". Undermining the assumption that cognition is information processing is undermining something fundamental to much of cognitive science, as it has traditionally been approached. And once that assumption has been undermined, both the problems offered by many of the traditional approaches to cognition, as well as their proposed solutions, appear to have been somewhat discredited.

Rather than viewing cognition in terms of a brain processing information, ecological, or "embodied" approaches (as I'll call them from now on) endorse the thesis that the unit of analysis for cognitive science should be the whole organism, situated in– and interacting with– a changing physical and social world. This approach often sees, not input to the brain, and then brain processes producing output, but a constant, complex, dynamically changing feedback-dependent system.

A detailed presentation of these "embodied" approaches, would take more space than I have here. (The references I referred to in the previous section are a good place to look for more detailed presentations.) My purpose is simply to give a more general overview of the methodological, epistemological and ontological outlook that these approaches share, so that I can then go on in the next chapter to address the issue of intentionality, and in subsequent chapters to show how such approaches could explain intentionality.

1.4.1 *Look at whole agents, not at brains.*

One central point is that a human being should not be bifurcated in the way that Cartesian dualists and most physicists do, into a purely thinking (i.e. not acting) thing –a brain– connected to and controlling a non-thinking body. In embodied theories, the fundamental locus of cognitive activity is not the brain. Rather the fundamental unit of analysis is a whole person. And a human being is seen, not as a thinking thing attached to a body, but principally as an *agent;* a whole indivisible agent, embodied in a certain way, and embedded in a certain physical and social context; an agent capable of performing rather complex cognitive actions. For instance, an agent capable of performing purposeful actions that aim at achieving certain goals, and capable of consciously entertaining those purposes and goals. And, importantly, a human being is an agent capable of performing linguistic actions (with higher-order levels of intentionality), and of interpreting the linguistic actions of others.

1.4.2 *Explain agents' actions and capacities*

This non-bifurcated view of a cognitive agent puts action, as opposed to thinking, at centre-stage. The focus here is more pragmatic than representational. The central phenomenon to explain is the actions of whole indivisible agents, rather than the thinking or representing or perceiving of a mind (brain). Thinking, representing and perceiving are all underpinnings of an agent's ability to act in certain ways. But —importantly— they are not separable from one another; and they are especially not separable from action.[22] The way agents experience their world(s), the way they think, and the ways they represent their world(s) are all intimately and inextricably intertwined with the way the agents act, the sorts of actions they are capable of performing and disposed to perform, and the physical and social contexts and situations in which the agents act.

1.4.3 *Mind is not separable from body and world*

From this embodied action perspective, mind-body-world is a package deal. An agent's mind, body and world are not separable, not even in principle. The ways I think, perceive, and act, and the world that an agent thinks about, perceives and acts within, all emerge *together*. The kind of thoughts I have depends upon the kind of body I have and what it can do, on the kind of world I live in and the actions it affords beings with bodies like mine, and on the social world that I act within, and the distinctions and practices that *we* make. Change my body and you change the way I think. Change my world, even my social world, and you change the way I think. So the best way to study my cognitive processes is to study, not my (conceptually isolated) brain and its inputs and outputs, but the whole system of brain-body-world, and the way it operates as a system. These

[22] In addition, veridical representation is not seen as all-important for embodied theories, as it is for many advocates of the representational theory of mind. Successful action is the phenomenon to explain, and to the extent that veridicality is useful for this end, it's important. But veridicality is not always the most success-producing option. Erring on the "safer" side (e.g. a rabbit's running away from things that might be foxes) is often more success-producing than strict veridicality (only running away when you're sure that the thing you see is a fox). Of course we are owed an account of what makes an action count as "successful" here. I give one in section 1.4.11

interdependent features of cognitive systems and the context in which they operate have been brought forth together, mutually influencing one another throughout the history of the human species, and throughout the learning history of each individual human being. Thus, the evolutionary forces that shaped our species and our physical environment, the cultural and artifactual environment we live in and the means by which it has adapted over time, and the learning and socialization that individuals undergo as they learn to participate in the practices of their community, are all relevant to understanding cognitive systems and how they are able to do what they do.

1.4.5 The cognitive system extends beyond the brain

On this approach, cognition is not simply a matter of brain processes. If we refer to our cognitive system as the apparatus that enables us to exercise our more complex capacities, then this system includes our bodies, and our physical and social worlds. Port and Van Gelder (1995b) put it like this:

> Since the nervous system, body and environment are all continuously evolving and simultaneously influencing one another, the cognitive system cannot be simply the encapsulated brain; rather, it is a single unified system embracing all three. The cognitive system does not interact with the body and the external world by means of periodic symbolic inputs and outputs; rather, inner and outer processes are coupled so that both sets of processes are continually influencing each other. (p. 13).

The cognitive *system* is best seen as the interaction of a set of interdependent entities, rather than the behavior of one entity (my brain), insulated from the external influences on it by sensory transduction and action. Much of our capacities depend upon embodied "know how"; knowing how to ride a bicycle, how to juggle, and how to hit a baseball are capacities that do not simply depend upon my having limbs with which to exercise these skills, but are capacities of which we would say that the capacity is itself "embodied". It is a bodily skill, and my body participates in learning, exercising, and in "knowing how" to exercise the skill. The knowledge is not only in my brain but in my body's "motor habits".

Our minds are composed of tools for thinking, but (as Dennett (1997) argues) we often leave these tools in the world rather than incorporate them into our neurological structures. Dennett also argues that humans' superior intelligence results not from our bigger brains, but primarily from

> "our habit of *off-loading* as much as possible of our cognitive tasks as possible into the environment itself—extruding our minds (that is, our mental projects and activities) into the surrounding world, where a host of peripheral devices we construct can store, process and re-represent our meanings, streamlining enhancing and protecting the processes of transformation that *are* our thinking." (Dennett 1996, p. 134-5)

There are many examples. One is the fact that many people use notebooks as ways to extend their memories and use pencil and paper to extend their calculating abilities. The system that instantiates my ability to perform long division includes the pencil and paper on which I calculate. Labelling kitchen canisters greatly simplifies memory tasks. Also, our worlds are often organized in such a way as to make cognitive tasks simpler, for example, numbered streets and avenues, and the Library of Congress catalogue system. John Haugeland (1995) tells about how Interstate 17 South is an integral part of the cognitive equipment that he uses to get to San Jose. He gets on the road, drives south and gets off at the end. By such innovations we drastically simplify the cognitive demand that living a human life places on any individual. Andy Clark reportedly says, "We make our world smart, so we can be dumb in peace".[23] In these and other ways the *world itself* is integral to our cognitive abilities. We would not be able to do what we do without these external "scaffolds" as Clark (1998, ch. 9) calls them.

But the world provides more than just an external memory and a place whose organization simplifies the cognitive demands of some tasks. The world is also a venue for performing operations that transform the cognitive problems posed to an agent. For instance, we often solve spatial orientation cognitive tasks, not by mental effort but by actually manipulating objects. Clark (1998, pp 65-6) reports research whose results show that advanced Tetris players rotate the shapes on the screen manually, to see its match with the geometrical opportunities the shape is falling towards. These players reduce the inner computational effort by externally rotating the shapes; apparently this is much faster and more reliable than imagining the zoid rotating. In addition, Clark's epilogue gives the example of writing a paper: it involves repeated interactions with writing on computer screens and pieces of paper; organizing, re-organizing, re-phrasing and adding to and subtracting from what's written in order to

[23] Dennett (1997) attributes this expression to Andy Clark.

produce the paper. Writing a paper is not (at least not by me) something done using one's bare brain: rather it involves repeated iterations of a complex feedback loop involving interaction between brain, hands, and external objects that are manipulated *in* the process of writing. Neurological structures are only part of the "equipment" used in such cognitive processes; the pieces of paper are also involved. The cognitive system includes the paper, computer, body and central nervous system.

This position in turn fits well with Ryle's (1949, p. 28 ff.) position, that much of the cognitive tasks that we can do "in our heads" are skills acquired slowly and with much effort. And these abilities depend on our first learning how to practice these skills in the more public realm. We learn to count, think, add, read, speak to ourselves, and theorize "in our heads" through first practicing these skills "in the world". It's through experience in embodied action, responding to and interacting with objects in the world, in the rather public performance of cognitive tasks that we learn many of our cognitive skills. And as these cognitive skills develop, we gradually learn how to "keep our thoughts to ourselves" as we exercise those skills.

1.4.6 A cognitive system can consist of many agents cooperating.

A cognitive system is often also a cooperative cognitive system. The cooperative teamwork of a group of people might bring forth a cognitive system that is best pointed to by pointing to the team, rather than to each individual team member. The achievement of a goal –winning the match, navigating the ship into port, winning the account, producing the film– may often best be described as a cooperative activity, where the locus of cognitive activity is in the interaction between the members of a group, rather than in each individual person. This idea is the dominant theme in Hutchins (1995) *Cognition in the Wild.*

1.4.7 Perception is for action, not for representation

One of the central tenets of this approach is that action and perception cannot be separated. This point is central to Varela, Thompson and Rosch (1991), who argue that perception and action are "fundamentally inseparable in lived cognition" (p. 173). They argue that perception consists in *perceptually guided action.* In a related fashion, Clark (1998, ch. 2) summarizes research showing that many of our perceptual processes are intimately bound to the actions those processes are used to guide. These studies suggest that the neurological structures activated by perceptual processes are not general-purpose, context-free representations (as many computational accounts would have it). Rather, these structures (whatever they are) implement special-purpose know-how, part of what dynamic systems theorists Thelen and Smith (1994) call *action loops*: neurological structures custom-designed to initiate and perceptually guide only certain types of actions. One classic example Clark cites (p. 37), is infants' slope-detection. The infants in one of Thelen and Smith's studies were placed at the top of slopes of varying steepness. Many of them unhesitatingly tried to crawl down, and often fell (they were caught in time). As they became experienced at crawling down slopes, however, they learned to recognise and to only crawl down slope gradients which were not too steep to crawl safely down. However, once they developed the capacity for walking, over two thirds of the new walkers unhesitatingly plunged down steep slopes, slopes they previously avoided crawling down. The know-how about slopes did not transfer to the new mode of action. They had to learn about slopes all over again when applying the perceptual discrimination abilities to the new mode of action.

It has long been argued (e.g. Piaget 1954) that a child's cognitive abilities (such as the ability to use perceptual information to guide action) develop through the child's experience in perceptually guided action. Clark concludes that in addition to this fact, much of the information infants learn is action-specific. Many of our perceptual processes are shaped through experience to detect features of the world relevant to guiding and initiating only certain types of actions, rather than detecting and representing features of the world used generally for all action. The introduction to Dreyfus's (1992, especially pp. xxvi-xxxiii) *What Computers Still Can't Do* also criticizes assumptions about the use of supposedly general-purpose "context-free" representations (as used in knowledge-engineering based AI). Dreyfus advocates thinking in terms of humans' context-embedded know-how, generated by experience of embodied action in relevantly similar contexts. Most of the time perception is for guiding specific types of action, not for representation per se.

Based in part on conclusions about the amount of human knowledge that is special-purpose know-how, and in part on conclusions about "action loops" like those mentioned above, neither perception nor action can be separated into two parts —the mental (neurological) part and the bodily part— as both Cartesians and physicalists do, and as is typical in cognitive science.

1.4.8 Mind and action cannot be treated separately.

And just as perception and action cannot be treated separately, mind and action cannot be treated separately. Much physicalist and AI research (not to mention Cartesian skepticism about other minds), is fuelled by the assumption that something could behave like a conscious being (one with a mind), when it is not in fact conscious. Objections to the Turing test as a demonstration of intelligent action are posited, the possibility of "zombies" is seriously entertained, and distinctions are invented (e.g. Searle's (1994) distinction between genuine and as-if intentionality) which presume that this is possible. The embodied action perspective I'm advocating rejects the assumption that something could fail to have a mind yet still behave as though it had a mind. (Thus if something didn't really have a mind, this would be apparent in the way it acts and interacts.) Intelligent (conscious, thoughtful) behaviour is the *criterion* for en-mindedness; en-mindedness is not simply one possible "cause" of such behaviour. To have a mind just *is* to be able to behave like this. It is a defensible criterion, as Dupré (1996, p. 330) argues; individual performances might be attributable to luck, or coincidence. But it does not make sense to claim of a being which *consistently* demonstrates a capacity to act in intelligent, thoughtful, conscious ways, or a being that demonstrates patterns of behaviour *only* explainable from the intentional stance, that it does not really have a mind or does not have "intrinsic" intentionality. If we view people as whole, embodied, context-embedded agents, then, as Ryle (1949) maintains, intelligence, purposefulness, consciousness, thoughtfulness are *in* the agent's behaviour (especially, for humans, in their participating in shared practices), not in some separate cause of the behaviour.

1.4.9 The world an agent acts within is the world-as-experienced by the agent.

The focus on the interdependence of perception and action also gives rise to the claim in embodied cognitive science, that beings with different capacities for perception and for action, in a very real sense "inhabit different worlds". Varela, Thompson and Rosch (1991) summarize the focus of embodied cognition like this:

> "[T]he overall concern of an enactive approach to perception is not to determine how some perceiver-independent world is to be recovered [through representing it correctly]; it is, rather, to determine the common principles of lawful linkages between sensory and motor systems that explain how action can be perceptually guided in a perceiver-dependent world." (p. 173)

Each creature acts within a "niche" of objects characterized with respect to their relevance for the creature's perceptual capacities, goals, dispositions and abilities. Clark (1998) puts this well when he remarks that "Biological cognition is highly selective, and it can sensitize an organism to whatever (often simple) parameters reliably specify states of affairs that matter to the specific life form." (p. 25). Ecological psychologists, such as J. J. Gibson (1979) and Ulric Neisser (1976), and biology of cognition theorists Maturana and Varela (1980) (see also Varela, Thompson and Rosch 1991) focus a lot on the way the world an agent perceives and acts within is an such agent-relative niche (the agent's "medium"), rather than an objectively characterized world. Braitenberg's (1984) *Vehicles,* argues for a similar perspectival limitation to the context within which an agent acts. The idea of niche-dependent perception isn't new. Clark (1998) refers to Von Uexkull's (1934) concept of an *umwelt*: the set of environmental features to which a given type of animal is sensitized. Different types of animals living in the same physical environment, can inhabit different "effective environments", to Von Uexkull. This gives rise, for example, to Gibson's (1979) notion of an object's *affordances,* the actions the object enables the agent to perform. What an object affords is a property of the object, but it is a property *for* the agent, and may not be so for a different agent with different abilities.

This agent-relative niche is "brought forth" *through* the agent's actions and abilities, over the agent's history of learning to act more successfully, and acquiring these skills and abilities. It is also brought forth through the agent's species' history of actions and interactions within an ecological niche (Varela, Thompson and Rosch 1991, pp. 201-2). Through evolutionary selection over many generations, selecting for dispositions and abilities to act more effectively, these interactions have shaped both the perceptual systems and bodies of individual members of the species, and also shaped the species' ecological niche. (The colour-vision of bees and flowers' ultraviolet reflectances are a good example of such co-evolution of a creature and its niche.)

1.4.10 Cognition depends upon shared practices.

In addition, the history of a human culture's actions and interactions establishes certain practices and conventions that in turn shape the distinctions people in the culture make.[24] These cultural distinctions and practices –also brought forth

[24] This is reminiscent (intentionally so, on my part) of Wittgenstein's (1958) talk about "agreement in judgements" as well as in definitions (§241-2).

through a history of (inter)action– are also part of the world a human agent experiences and acts within. As they develop and grow, human agents are encouraged into conformity with these practices, a process that ensures a large amount of intersubjective agreement in judgements (distinctions) about what is seen (equivalent to judgements about what is there, for them) among different individuals within the cultural group. Their existence brings about what Neisser (1989) calls cultural affordances,[25] such as the fact that to me a mail-box affords letter-mailing; the existence of practice of putting letters into mailboxes and the accompanying practices of people taking mail from the box and eventually delivering it to the address on the envelope make it the case that the mailbox affords letter-mailing. Because of this practice and the role that mail-boxes play in the practice, I *see* the mailbox as something that affords letter-mailing.

The practices that exist within a culture support and sustain many human cognitive capacities. This theme is prevalent in Haugeland (1995), Brandom (1994), Dreyfus (1991), and to a lesser extent in Dennett (1995a). In many cases our cognitive capacities cannot be understood outside of the network of interlocking practices that support and are supported by human cognition, as it is exhibited in people's interactions. My ability to get a door closed by asking you to close it is a simple example. My asking makes little sense unless we see it as situated within a shared form of life that includes the practice of asking people to do things and doing things for people when they ask you to. My ability to write this dissertation depends upon many cultural "props". For instance it is set within the practice of people producing ideas and publishing their ideas for others to critique and employ (itself a cooperative cognitive activity, instituting a "community of inquiry") and the practice of writing and defending dissertations. It depends upon people being required to have Ph.Ds to teach philosophy, upon the practice of employing graduate students to teach courses so that graduate students can eat while writing, upon the practice of growing, shipping, selling and buying the food that sustains me as I write, and upon the practice of renting dwellings that enables me to have a place to write. My ability to write this dissertation also depends upon the practice of writing using a word-processing program on a computer, and on the practice of producing and maintaining such

[25] This is an unpublished conference presentation. I'm using Neisser's distinction between physical affordances and cultural affordances as cited by William Bechtel (1990, Bechtel and Abrahamsen 1991). For example a book culturally affords reading, while it also physically affords propping up wobbly table-legs, swatting flies and weighing down pieces of paper so they don't blow away.

programs and computers. (In other times it depended on the practice of using pen and ink and paper to write on, and on the practice of producing pens, ink and paper for people to write with and on). These, and thousands of other large- and small- scale practices like them, are the conditions for my ability to produce a dissertations such as this. And similarly for almost every human action: almost everything we do beyond basic biological processes like breathing and walking and urinating is set within an interwoven context of practices; and even these basic biological processes are subject to norms governing where and when we walk and urinate and what we breathe. The presence of, and conformity with, these norms structures our lives in uncountable subtle and not so subtle ways.

Later on I am going to be making a big deal out of the fact that an explanation of most human capacities depends upon an explanation of the practices that are the preconditions for the possibility both of that capacity coming into being, and of its exercise. That these practices, and the criteria by which a performance counts as conforming with a particular practice, are *shared* by members of the community that practices them is also important.

1.4.11 Purposeful actions have a special kind of setting, not a special kind of cause.

A person's action does not break into two parts, the mental volition preceding a bodily action. Here (as per Ryle's (1949) category mistake argument) "mind" is not used to refer to a separate process or arena in which other "hidden" events (such as volitions) in addition to the publicly observable bodily movements take place. Ryle argues that a bird's migratory actions don't break into two parts: flying south and migrating. Rather, the bird is flying south "migratingly". "Migrating" refers to the *way* the bird is flying south. It involves considerations of the context of the bird's flying: the bird's destination, motivation, abilities, the time of year, and perhaps whether other birds of that type are also flying south. Similarly, acting purposefully does not break down into two parts: the mental (neurological) intending or willing, and the physical moving. A purposeful action does not merit the appellation "purposeful" because it is caused by a special kind of event: a mental one. Rather, it merits the being called "purposeful" by virtue of the context of the acting. What makes an agent's actions those of an en-minded agent is not the agent having a special thinking part, in which precursors or accompaniments to the bodily action takes place. What makes the agent's actions those of an en-minded agent –what makes these actions count as *purposeful* actions– is partly the physical and social context (for example, in response to what they were done, in what manner they were done) and partly the agent's other abilities and actions (the agent's inclination to take responsibility for it, the agent's ability to repeat the performance, to perform related tasks, to answer questions about it, etc.).[26]

"Having a mind" is equivalent to "being able to act in certain 'intelligent' ways", being able to slope arms *obediently,* to answer *thoughtfully,* to argue *rationally,* to thump my fist on the table *intentionally*. This "is equivalent to" does not signal a reduction. Rather it signals the kinds of things we take as criteria for the expression "has a mind" to be used as it should be used.

There are not two parts to my action of intentionally thumping my fist on the table: intending and thumping. Rather my thumping the table intentionally, refers to the *way* I did it: intentionally (or perhaps deliberately), as opposed to accidentally, automatically, absent-mindedly or under duress. (See Ryle's remarks in Bestor (1979) about this "adverbial" approach to the mind.) This "way" I perform the action, that determines whether my action counts as intentionally thumping my fist on the table, refers to the many physical and social criteria[27] the action must meet for it to count as intentional. (I refer here to whether my action "count as", as opposed to "really is", intentional, very deliberately; these are defeasible criteria for the appropriate use of the word "intentional", not for whether the action *really* was intentional.) These criteria can include whether it was expected that I would do this (especially whether *I expected that I would do it*); whether I look and/or feel satisfied, as opposed to surprised or reluctant, when it happens; whether I am attending to what I am doing, or so engaged in the argument that I'm unaware of the gestures that I'm absent-mindedly making. It also depends on counterfactual future actions; for example, upon my willingness to take responsibility for the consequences of performing the action. If I break the table or scare the cat, others could hold me accountable for such consequences, and I should accept that responsibility. The

[26] Admittedly, it does make sense to sometimes speak of the willing side of an action, as a separate activity from the moving; for example if my leg becomes paralyzed, I could say that I'm willing my leg to move, but it isn't moving. That fact that we can sometimes say this in such unusual cases, does not mean, however, that in everyday cases such a bifurcation of the action is a sensible move. This kind of tactic is very prevalent in the writings of Wittgenstein and Ryle.

[27] Note that I did not say "mental" criteria here. I did this deliberately. This is partly due to the fact that what counts as my being in a certain mental state (e.g. intending to thump my fist on the table) itself depends on physical and social criteria. It's also, as we'll see later, because of the fact that in order to play a role in our interactions, as Wittgenstein (1958) argues, "An 'inner process' stands in need of outward criteria" (§580)

action, if intentional, also commits me to certain preconditions and consequences of having done that: it commits me to the position that it was justified, for instance, and to the position that I should give reasons for doing so if they are asked for.[28] (I will be paying a lot of attention to these social practices, and the criteria of conformity employed in them, in the next chapter.)

1.4.12 Perception does not (always) involve representation

Just as action should not be seen as a two-part process, perception should not be treated as a two part process: the physical processes in the eye and optic nerve, and so on, causing a mental representation to be activated by these processes (a representation used to guide actions involving the object(s) represented). The traditional claim, recall, characterizes a cognitive process as an "information processing" or "computational" process operating on the formal or structural properties of a representation. Thus sense-organs are thought of as transducers, and the "outputs" of the transducers are thought of as "representing" the external stimulus that caused these sense-organs' outputs. One of the more hotly contested debates in current cognitive science is that of the role of representations in explanations of cognitive phenomena. Some theorists criticize this traditional view, arguing that representations, as they have traditionally been conceived, *never* play a role in cognition. Others counter with a middle-ground position, arguing that representations are *sometimes* involved in cognitive processes, but only in certain cases, and that claims about such cases have been over-generalized when applied to claims about all cognition.

Since I am going to be arguing for a view on intentionality, and how we can productively view the problems associated with naturalizing intentionality, from an embodied cognition perspective, I'll need to spend a little time outlining this debate, and clearing up the terminology in this debate. That's the task I take on at the beginning of the next chapter. I use this to show the important role of normativity in explanations of intentionality.

[28] Brandom (1994, esp Chapter 3) focuses on the practice of "deontic scorekeeping" we engage in. We keep track of how people's actions (including our own) commit the actor to performing other actions, accepting certain positions, undertaking certain responsibilities. This practice of scorekeeping and the practice of giving and asking for reasons that it is set within, for Brandom, is one of the fundamental shared practices that underlie much of humans' capacities.

Chapter Two

Representation in Cognitive Science

If the only tool you have is a hammer, you tend to treat everything as if it were a nail.

– Abraham Maslow[29]

2.1 Representations in Embodied Cognitive Science

As I have been saying, computational cognitive scientists argue that representations *always* mediate between perception and action. "Cognition can best be understood as information processing" is a claim typical of texts in cognitive science.[30] And within such a tradition, the "information" is usually thought of as *represented* information. For example, Barbara Von Eckardt (1993, p. 50) argues that most cognitive scientists (at least tacitly) accept the assumption that the human cognitive system is a representational and computational device, and that

> A computer is a device capable of automatically inputting, storing, manipulating, and outputting information in virtue of inputting, storing, manipulating, and outputting representations of that information....(p. 50)

Thus the mind/brain's capacities, on this view, can be thought of along the lines of something that takes represented information as input and produces further representations of information as output. These input representations come from sensory transducers constructing representations of ambient stimuli, or from memory. The outputs are either representations of actions for the body to carry out, or representations of information to be stored in memory. These assumptions "give cognitive science its identity," says Von Eckardt (p. 50). There are disagreements among cognitive scientists, of course, about the nature of the representations (explicit or tacit, language-like or picture-like, atomic or distributed, etc.). There are also disagreements about the nature of the

[29] Cited by Deacon (1997), p. 47.

[30] For example, see Von Eckardt (1993, p. 50), Dawson (1998, p. 5)

computational processes (like a digital computer processing symbols or like a connectionist network activating nodes on the basis of patterns of input node activation). These disagreements are in house disputes, however, set within a large amount of agreement. Dawson (1998) argues that, in spite of all these disputes and disagreements, and in spite of the fact that cognitive scientists come from many different disciplines, sharing the assumption that *cognition is information processing* unites cognitive scientists into a community bonded by a common language (pp. 6-7). Sharing this assumption enables different cognitive scientists to talk to one another, and to understand what one another's positions are in the various debates within cognitive science.

Many embodied action theorists oppose this assumption that cognition is representational information processing. They argue that representations rarely (some say never) mediate between perception and action. The basic criticism of claims that representations are *always* involved in cognition, is that in many cases, it is quite counterproductive to think of perception in such bifurcated representational terms. A lot of very productive work, they argue, has been accomplished without thinking of cognition in this way. They argue that rather than constructing or activating a representation of the objects perceived, and then acting on the basis of the represented information, we often respond directly to the objects as *presented* in perceptually guided action, not as *represented* in the mind (brain) as a result of perceptual processes. Such theorists defer to examples such as Rodney Brooks' (e.g. 1991a, 1991b, Brooks and Stein 1993) autonomous robots, which he says do not construct representations of the world they operate in. Instead, the robots employ feature-detectors that affect the robot's actions only when those features are detected. For example, a robot employs a wander mode that causes it to wander randomly (exploring), unless a short-distance obstacle detector (e.g. a radar system, detecting the presence of an object in the robot's path) is triggered. The activation of this detector subsumes control from the lower-level "wander" module, causing the robot to stop moving forward, and thus it avoids crashing into the obstacle. Brooks (1991b) argues that this system does not involve *representing* the obstacle and reacting to the representation of it, but simply *detecting* the obstacle and reacting to the obstacle itself. Brooks concludes:

> When we examine very simple level intelligence we find that explicit representations and models of the world simply get it the way. It turns out to be better to use the world as its own model (p. 140).

Such examples suggest that an agent is (at least often, for very "simple level" tasks) best characterized as responding *directly* to features of the agent's environment, rather than responding *indirectly*, via responding to internal representations of those features. ("Classical" cognitive scientists (e.g. Vera and Simon 1993, p. 33-5) object to this claim that there is no representation involved here. The output of the feature detector, they protest, is a representation (symbol) of the presence of the feature detected. Obviously, the notion of what counts as a representation needs to be cleaned up. I'll get to that presently.)

Many theorists (dynamic systems theorists in particular[31]) take such examples and such conclusions very seriously, and use them to deny that representations are *ever* useful entities to postulate in explaining cognitive phenomena. When observing a system (machine, insect, animal, person) we may be tempted to explain its behaviour in terms of internal representational states and processes. However, these theorists argue, such behaviour can be explained without the need to invoke representational states—states that *stand for* something else. Explaining cognition in representational terms, they argue, creates a misleading picture of cognitive mechanisms, many –perhaps most, they argue– of which are better explained in terms of dynamic interactive feedback-controlled systems.

The classic example in dynamic systems theory is the Watt governor. Tim van Gelder (1995, p. 347 ff.) is often credited with making this analogy. However, back in 1976, Richard Dawkins (1976) introduced the same example to make a similar point.[32] Dawkins uses the Watt governor as an analogy to the way that it can be tempting to attribute purposiveness, even conscious desires and goals, to purposeless low-level biological processes (Dawkins talks about "an animal 'searching' for food, for a mate, for a lost child"). These can *appear* to be representational and purposive, but they have simple non-representational non-purposive explanations. The Watt governor is a mechanism designed to ensure that a steam engine's flywheel rotates at a continuous speed, in spite of

[31] See Port and van Gelder (1995a). For a general introduction to this way of viewing complex systems, see Abraham and Shaw (1992).

[32] As reprinted in Hofstadter and Dennett (1981), this is on p. 134-5.

fluctuations in the pressure of the steam driving the flywheel (due to changes either in the workload of the engine or the heating of the boiler) by virtue of a valve being slightly opened or closed to let more or less steam out. Looking at the behaviour of the system, it would be sensible to presume that since the flywheel returns quickly to a constant speed, a controller of some sort must be monitoring the flywheel and controlling the steam valve. Such a controller must be consulting a representation of the current flywheel speed, comparing it to the desired speed, calculating the appropriate amount to open or close the throttle-valve to adjust the flywheel's speed back to the desired speed, and then adjusting the throttle-valve by that amount. Such a representational explanation would account for the behaviour of the system.

This is not how the system works, however. The Watt governor is not a system that represents the speed and adjusts the valve according to the represented speed. Rather, the Watt governor is a set of arms mounted on a spindle geared into the flywheel, such that as the flywheel accelerates its rotational speed, the rotational speed of the spindle increases proportionately. On the end of each arm is a metal ball. The arms are hinged so that as the balls' rotational speed increases with the increasing speed of the spindle, the balls' rotational inertia causes the hinged arms to move outwards. These arms are connected by an ingenious mechanism to the throttle valve, which controls the steam pressure that drives the flywheel. The arrangement is such that when the flywheel speeds up (due to changes in the steam pressure or workload), the spindle speeds up as well and the arms move outwards. This causes the valve to close slightly, decreasing the steam pressure driving the flywheel and so slowing the flywheel's rotation. And as the flywheel slows down, the arms move back inwards due to decreased rotational inertia, causing the valve to open slightly, increasing the steam pressure and thus increasing the flywheel's speed. This system keeps the flywheel of the engine rotating at a constant speed, almost instantly compensating for fluctuations in the pressure of the steam and the flywheel speed. The point of this example is that nowhere in the system is there a representation of the steam pressure, nor of the rotational speed of the engine. The whole system must be understood in terms of the continuous reciprocal causal interaction of the parts of the system, consisting of the steam pressure, flywheel, spindle, weighted arms and throttle valve.

In the Watt governor nothing *stands for* the speed of the flywheel. The rotational speed of the spindle is caused by the speed of the flywheel, and so someone might want to say that the rotational speed of the spindle is used by the system as a representation of the rotational speed of the flywheel. But this is only because the driving force is on the flywheel, and we *want* to separate the spindle and the flywheel it is geared into and to say that the flywheel is the primary thing turning, and the spindle (or the balls) is *caused* to turn by the flywheel. Dynamic systems theorists could argue that a separation between the flywheel and the spindle (or the flywheel and the rotating balls) is a rather arbitrary place to separate the components of the system. Let's say that the force from the steam-pressure is exerted at the circumference of the flywheel, and the spindle geared into the centre of the flywheel. If this were the case, we could be just as justified (because just as arbitrary) in separating the outer rim of the flywheel from the centre of the flywheel, and say that the turning at the rim causes the centre to turn. Comparisons like this make it seem rather arbitrary to "cut" the system at one place between cause and effect, and to say that one part causes the other and the caused part is used by the rest of the system as a representation of the cause. It could just as easily be argued that the governor system is one-of-a-piece, and changes in steam-pressure cause (via the mechanisms I have just described) compensating adjustments in the steam pressure. No representational talk is needed to describe this process.

The Watt governor is used as an example of the kind of causal, non-representational system that gives rise to much of a cognitive system's behaviour. Clark (1998, p. 171-2) illustrates the idea that a similar kind of "continuous reciprocal causation" is also involved in people's cognitive activities, using Merleau-Ponty's (1942, p. 13) example of trying to catch a hamster with a set of tongs as it runs about on a table surface. Here it makes sense to see me as responding *directly* to the hamster itself, rather than responding to it *indirectly* via responding to a representation of it. As I move my hands and thus the tongs in my hands to try to catch the moving hamster, I respond to perturbations of my visual system. But this visual stimulation requires me actively to move my eyes and head in response to the stimulation I receive, so that I can continue to focus on the movements of the hamster. Furthermore, I move the tongs in response to the movements of the hamster while the hamster's movements are a response to the movements of the tongs. The important phenomena here are a kind of iterated interactive "dance" that involves a whole system constituted by the hamster, tongs, the table surface, and myself. This system does include cognitive, perceptual and sensorimotor aspects of my (and the hamster's) neurological apparatuses. But the basic locus of cognitive scientific explanation and interest is not on the *isolated* (or perhaps insulated) brain mechanisms and their inputs and outputs. Rather the whole dynamic system, including the neurological aspects, is the focus of analysis and explanation. This claim is motivated by the belief that to isolate my neurological mechanisms, and to treat the rest of the system as "external", in the sense that it could be replaced by a series of inputs and outputs, would be to focus away from a very important aspect of the phenomenon under investigation. This aspect is the whole cognitive system, and its behaviour. The cognitive system in this case, however, is not limited to my brain; it encompasses the whole interdependent system. As Clark points out (1988, p. 163), when analyzing phenomena like this, positing boundaries at the sensory and neuromuscular components of my body begins to appear to be positing rather arbitrary parts of the system to "cut" it at. Neither the hamster nor I enjoy any special status in explaining the behaviour of the system we constitute. The boundary could just as easily be drawn at my wrist and at the edges of the tongs. Thus, we could focus on the tongs and my hand, treating the hamster and the rest of me as sources of perturbations of that aspect of the system. To understand the behaviour of the whole system, we need to keep the continuous reciprocal causation at play between *all* the elements of the system firmly in view.

Many embodied action theorists argue that much of our cognitive processes are best described, not in terms of computational processes operating on representations, but in terms of non-computational and non-representational concepts and explanatory schemes used to explain the behaviour of complex dynamical systems. So for example, the tools of dynamical systems theory are used to explain cognitive processes in terms of point attractors, basins, and vectors in multi-dimensional state-spaces. Some even argue that *all* of human cognition can be explained this way. This claim is rather strong. However, versions of it can be found in recent work in developmental psychology (Thelen and Smith 1994), in robotics (Brooks 1991b, although Brooks' claims are weaker than they are often interpreted to be), in embodied cognitive science (Varela, Thompson and Rosch 1991), and in neuroscience (Skarda and Freeman 1987).

According to such theorists, all of human cognition can best be explained in non-representational, non-computational terms.

Port and van Gelder (1995a, p. 31), however, paint this as a methodological, rather than an ontological, assumption. They argue that the dispute between computational/ representational cognitive science and dynamical systems theory is a border dispute. Computationalists try to account for *all* cognition in terms of their concepts and tools, keeping all cognition within the representational domain. Dynamic systems theorists draw the boundary to include all cognitive processes within the dynamical domain. This claim, however, is not a pronouncement about how things really are, but rather a guiding methodological assumption. "It remains to be seen to what extent this is true," they argue (p. 31), "but dynamicists in cognitive science are busily attempting to extend the boundary as far as possible, tackling problems that were previously assumed to lie squarely in the computational purview." Thus it is largely an empirical question whether non-representational explanations can be given for all cognition.

Some theorists argue that it is perhaps not an empirical question. They argue that dynamic systems theorists rightly take *some* cognition to be best described in non-representational terms, but argue that they over-generalize when they claim that all cognition can be explained non-representationally. There are some problems that cannot be removed from the purview of computational/ representational explanation. Clark (1998, p. 166 ff., Clark and Toribio 1994, p. 418-20) for instance, takes non-representational accounts to be acceptable, but only for *some* cognitive processes. Many cognitive processes –particularly our abilities to cope with immediately present objects and situations– may best be described non-representationally. My abilities to catch a fly-ball (see Clark 1998, p. 27), to ride a bicycle, to negotiate my way through a crowded marketplace, and to catch a hamster with tongs are possibly best explained in non-representational terms. That is, they can be explained in terms of a causal system *directly* interacting with features of the situation that are presented to it perceptually, instead of responding *indirectly* to them, via responding to the way that are represented.

However, some of our cognitive capacities, argues Clark, are not best explained in such terms. The exceptions are those cognitive phenomena dealing with what Clark dubs "representation-hungry" problem domains. One type of representation-hungry problem domain is situations that involve reasoning about environmental features that aren't (or aren't reliably) presented perceptually. In such cases that agent deals with "absent, non-existent, or counterfactual states of affairs" (Clark and Toribio 1994, p. 419). For example, thinking about or identifying the location of an object that is not currently present: trying to remember where I left my coffee cup, or reporting that Joe has gone to the store, for example. I can recognize Joe when he's here, but I can also think about Joe when he's not here. These abilities seem to require representational capacities, to enable the agent to think about objects that are not currently presented perceptually. Another type of representation-hungry problem domain involves situations where the agent must selectively attend to "parameters whose ambient physical manifestations are complex and unruly" (*Ibid,* p. 419). For example, sorting objects that are identified by virtue of an abstract property or by an open-ended disjunction of features, such as the task involved in identifying all the valuable items on the table, or all the items belonging to the Pope (*Ibid,* p. 420).

Most of the anti-representationalist advocates, Clark and Toribio argue, cite non-representation-hungry examples to further their cause, where "suitable ambient environmental stimuli exist and can be pressed into service in place of internal representations" (*Ibid,* p. 418). They then generalize explanations that serve these cases well, to posit that *all* cognition can be accounted for without the need for explanations in terms of representations. For example, Clark and Toribio cite Skarda and Freeman's (1987) conclusions from their "beautiful and challenging Dynamic systems model of the way sensory information is registered in the olfactory bulb" (Clark and Toribio 1994, p. 421) that go far beyond the conclusions their model licenses. Their model is not of a "representation hungry" problem domain, but they draw very general conclusions:

> The concept of 'representation' ... is unnecessary as a keystone for explaining the brain and behaviour [because] the dynamics of basins and attractors can suffice to account for behaviour without recourse to mechanisms for symbol storage (Skarda and Freeman 1987, p. 184)[33]

[33] As cited in Clark and Toribio (1994, p. 421).

Because such claims are generalized from non-representation-hungry examples, their generalization to representation-hungry problem cases is rather hasty. Non-representational explanations of cases where the agent responds to "simple physical properties detectable in the ambient input" (Clark and Toribio 1994, p. 422) do not refute representationalism. Although explaining these cases shows that the tools of dynamic systems theory are powerful and useful, and could greatly aid in understanding some aspects of cognition, such cases are not the cases on which anti-representationalists should rest their arguments. Clark and Toribio do not argue that the tools of dynamic systems theory cannot explain representation-hungry problem cases, however (at best they can offer a new way of understanding representations). They simply argue that it hasn't yet been shown that they can. As Port and van Gelder argue, however, it might be hasty to conclude that they cannot. This is, to a very large extent, an empirical question; the proof of the pudding will be in the tasting. Dynamical systems people assume that this can be done, and are trying to do it. We will perhaps have to wait to see whether they succeed.

Steve Torrance (1999) has a slightly different take on this debate. Torrance makes a distinction between representation-rich cognition, and representation-economical cognition. Torrance argues that even the reactive, world-embedded, embodied cognitive skills exhibited by a jazz pianist are to some extent representationally mediated; "...grabbed chords are not *just* grabbed—they are picked out on the basis of quite definite criteria, which latter are painfully acquired by beginning players" (p. 59). This last comment echoes Dreyfus and Dreyfus (1982) claims about how skills are acquired, at first by explicitly representing instructions, but this gives way to an especially "tuned" sense of perception of the moves demanded by the situation. Thus Torrance argues for a spectrum of more-or-less representational cognition, where situated, responsive, world-based action and internal-model mediated action are two extreme regions on a continuum (p. 60).

I think Clark and Torrance are probably correct. However, some cognitive processes seem to depend upon the agent's system being sensitive to factors not directly presented in the ambient stimulus. Whether the tools and concepts of dynamic systems theory can accommodate these capacities is probably something we will have to wait and see about. The attempt to formulate such explanations is certainly worth while.

However, this is not *entirely* an empirical question. Lurking in the background of this debate is a conceptual question that, left unanswered, seems to leave the participants in the debate about the extent to which human cognition is representational talking past one another.

2.2 *What Kinds of Things are Representations?*

This conceptual question involves the looseness of the term "representation". It is used in rather different ways by representationalists and anti-representationalists. Underlying the computational cognitive scientists' argument is the view that *any* internal state that is caused by an external event can be a representation of that event. (The only condition seems to be that it is, or can be, *used* as a representation by some process.[34]) Thus claims that *all* cognition is representational, is prima facie true in the sense that it all –even the "continuous reciprocal causation" examples like the hamster and tongs– involves internal neurological mechanisms that are in the states they are in because (at some perhaps far-distant point in the causal chain) they were caused to be that way by events external to the agent. Vera and Simon (1993), for instance, argue that when non-representationalists describe activities in terms of directly responding to the world, the perceptual processes involved in these activities are nonetheless representational ("symbolic", in their terms). For example, they argue that Brooks' robots "are very good examples of orthodox symbol systems: sensory information is converted to symbols which are then processed and evaluated in order to determine the appropriate motor symbols that lead to behavior" (p. 34).

To anti-representationalist cognitive scientists, however, such cases do not involve representations. To them, "representation" has a much more restricted use. To anti-representationalists, representations are internal states that do the job of *standing in for* external events or items, such that they can guide the system's behaviour in the absence of that external event or item. The kinds of dynamical systems I have described do not use anything that fits their definition of a representation; no *standing in for* needs to occur in the system to explain the system's behaviour. Their definition of "representation" is something like the

[34] This is explicitly stated by Von Eckardt (1993), p. 51. Millikan's (1989) stress on the "representation consumer" also reinforces this view.

definition offered by Haugeland (1991, p. 62), who depicts a system as representational just in case:

(1) By using a signal, it coordinates its behaviour with environmental features that are not always reliably present to the system.

(2) It copes with cases where the environmental features are not present, by having something else, other than the signal from the environment, stand in for the signal, and uses this to guide its behaviour.

(3) The "something else" is part of a general representational scheme which allows the "standing in for" to occur systematically, and allows for a variety of related representational states.[35]

Here (1) rules out as "non-representational" cases such as those the dynamic systems theorists refer to, where the agent is coordinating its behaviour with the features of the environment (the hamster, baseball, bicycle), via a "signal": the light rays from the object stimulating the agent's retinas, etc. As dynamicists claim, such cases are not representational (in this sense); nothing is used to *stand for* an object that may or may not be present. However, such cases are still representational in the sense employed by representational/computational cognitive scientists, in that the environmental features do causally influence the system's internal states. These internal states are employed by the agent's cognitive system to cause behaviour that is coordinate with or directed towards the environmental feature that is the cause of the "representation".

Point (2) makes it clear that in this more restricted sense of "representation", only cases where something is used to stand in for the environmental features, such that they can be used to guide behaviour in cases *where the environmental feature is absent* (or even non-existent), count as representations. (Put in other words, only systems that can misrepresent –represent the presence of something that is absent– count as representational systems.) Internal states that enable an agent to respond directly to environmental features that are present to the agent's perceptual systems, do not necessarily represent.[36]

[35] This is originally from Haugeland (1991, p. 62), as described in Clark and Toribio (1994, p. 404) and in Clark (1998, p. 144).

[36] Motor emulators are an interesting exception raised by Andy Clark (1998, p. 22-3). Here an internal system models the position of limbs as the agent directly interacts with objects presented to it. Nerve signals sent to a limb are also sent to the emulator, which models the position of the limb, and sends back a signal that ought to be identical to the proprioceptive signal that indicates where the actual position of the limb is. Such

Point (3) talks about the kinds of systems that are *capable* of misrepresenting. The standard (but by no means universal) [37] wisdom in theories of content is that only entities that can have many different intentional states, can have any intentional states at all. Representations are part of a representational system. The relationships between the different elements of the system enable referential relations to be in place, even when the referent does not exist. For example, even though centaurs do not exist, the related concepts of "horse's legs", "human torso", and so on do have accepted referents. These related concepts "anchor" the referent of "centaur" such that, even if no such creatures in fact exist, we agree upon the criteria something would have to meet in order for someone referring to it using "centaur" to qualify as a correct or proper use of the term.

This account of what counts as a representation accords with the general requirement that only neurological entities that can *misrepresent* ought to count as representations (see Fodor 1990, Sterelny 1990). We want it to be the case that *some* things which *can* cause the internal state, are things that it does not correctly represent. Put another way, the representation has to be able to stand for the object it stands for, in the absence of that object (i.e. when it is caused by something other than the kind of thing that it correctly represents).

This also fits with the thesis that a cognitive entity that is able to think about (act with respect to) things that are not directly present in its environment has a distinct advantage. A creature that has the capacity to remember past events, plan future events, anticipate possible contingencies, and postulate explanatory theories, would have a terrific advantage over creatures limited to dealing only with what they are presently encountering. The drawback with the capacity to do all this, however, is the possibility that one can be wrong; the

signals get to and from to the motor emulator quicker than signals to and from the limb (they don't have to go so far). Using the motor emulator (and assuming that the limbs indeed go where they are being "told" to go) enables the system to quickly compensate for errors (differences between where the limb is and where it needs to be), and so to deal with a fast-changing environment, where speed and smoothness of response is important. According to Clark, such devices are widely employed in industrial control systems. This example shows a case where a representational system can be useful when interacting directly with objects that are directly present to the system. The point, however, is that except for such representational mechanisms that improve efficiency, they do not need to be used. They improve efficiency, but are not a condition on the possibility of the action happening at all.

[37] See, for example, Fodor (1990, p. 51). Fodor describes this condition as "the conventional wisdom," but doesn't endorse it (p. 52), since it implies the rejection of his atomistic view.

memory is incorrect, the plan doesn't come to fruition, the possibility doesn't occur, the explanation is incorrect. This capacity to get things wrong spills over into the capacity to misidentify objects perceptually encountered as well, so that something that is directly presented to the system is identified as something other than what it is (a possum-on-a-dark-night is represented as a cat, a stranger-at-a-distance is thought to be a familiar friend). Giving a naturalistic answer to the question of what gives rise to the ability to represent and to misrepresent, is the problem of naturalizing intentionality. This is the problem I'm taking on in this work.

I'm going to come back to this restricted sense of representation soon, and show that it echoes a similar distinction between types of representation made late last century. But first let's look at the kind of problem, and the kind of intractable disputes, that arise in accounts of representational content if we don't distinguish types of representation.

2.3 *Theories of Content.*

If we look at the standard literature on the topic, it would appear that the ability to misrepresent is a difficult phenomenon to give a naturalistic explanation of. Fodor's (e.g. 1990) *disjunction problem* points out the difficulty in naturalistically specifying the content of a representation in such a way it misrepresents when it is caused by something other than that type of thing. Naturalistic explanations of content, on Fodor's view, have to be able to distinguish Type One cases where the representation is caused to be activated by something that it *correctly* represents from Type Two cases where it is caused to be activated by something that it *misrepresents* (Fodor 1990, p. 60). The problem is that it is difficult to give a principled distinction between the two types of cases.

Fodor (e.g. 1990, p. 57-60) sees the activation of a representation in terms of a causal regularity. The activation of a representation here is seen along the lines of a causal law of nature. We should, he says, be after a description of the properties that are "nomically sufficient" (*Ibid.* p. 59) for causing the representation to be activated. The problem, for him, is that we can easily (in fact we *should*) state the rule that describes the regularity, such that the purportedly aberrant example (the bee-bee pellet causing the frog to snap, for instance) also accords with the rule. If these are causal laws, then since causal laws are counterfactual-supporting, the fact that the bee-bee pellet *can* cause the representation to be activated means that there is no principled non-intentional (i.e. scientific) justification for calling this a case of misrepresentation. The frog's perceptual state can be described as activating whenever a fly is present, but we could just as justifiably (perhaps even *more* justifiably) describe it as activating whenever a fly or a bee-bee pellet (or a "little ambient black thing"; *Ibid,* p. 72) is present. This is what happens, after all. Teleofunctional, causal-historical, informational and other theories of content cannot give a naturalistic account of why we should describe the representation as representing flies, as opposed to small ambient black things. Thus, it can be said to correctly represent *fly-or-bee-bee* or *small-ambient-black-thing* rather than just *fly*. Thus the frog's perceptual state does not misrepresent when caused to be activated by a bee-bee pellet.

In order to account for cases of misrepresentation, we need to have reason not to view the phenomena simply in terms of a causal regularity that *objectively* describes what happens. In accounting for a causal regularity (a causal law, if you will), the observer's task is to find the description of the regularity that best accounts for *all* the cases where the phenomenon occurs. The trouble is that most accounts of naturalized content fail to get beyond causal-law describable behaviour into accounting for rule-following (norm-governed) behaviour. Either that or accounts that attempt to naturalize content sneak in normative standards for what the representation ought to represent, and so are accused of failing to naturalize the content. This is the basis of Fodor's critiques of the many attempts to naturalize representational content. If it's a causal regularity, then *all* the things that can cause the representation to be activated belong in the representation's content, and so you cannot account for misrepresentation.[38]

[38] Fodor tries for a nomological (causal law based) explanation for why certain objects cause certain representations to be activated, yet a naturalizable (causal law-based) distinction between cases that count as cases of misrepresentation. Fodor's asymmetrical dependence theory –where non-cow-caused tokenings of COW are dependent on cow-caused tokenings of COW, but not vice versa– may non-normatively account for a difference between the two types of cases. A detailed response to Fodor's theory is beyond the scope of this work. However, as a first pass, it seems to me that he fails to give any account of what makes a certain representation something that represents at all, and what makes it a representation that represents cows such that the non-cow-caused activations are the aberrant cases. (Why isn't it a NON-COW representation, such that the asymmetry runs in the opposite direction, and the cow-caused activations are misrepresentations?). The answer, I think, has to be given in terms of the way that COW should (in a normative sense) represent cows.

But intentionality is not a causal-law based phenomenon; it is a norm-based phenomenon. Reminiscent of Hume's (1739) critique of the attempt to derive an "ought" from an "is", attempts to naturalize content can be seen as attempts to define a normative concept in non-normative vocabulary. In Fodor's terms, the representation *ought* to represent the things that caused it in Type One cases and not the things that caused it in Type Two cases. I don't think this formulation, in normative terms, can be escaped by re-wording. Intentionality *just is* a normative concept. As I will soon show, the notion of representation has normativity built into it at its very foundation. The attempt to account for content non-normatively[39] must have an answer to Hume-style problems of justifying a norm using non-normative vocabulary. Fodor's disjunction problem critiques are similar; basically showing that regularities in nature do not have exceptions, and that appeal to purely natural phenomena cannot justify describing some of the cases where a representation is activated as exceptions. This kind of attempt at naturalistic justification of a norm can be said to aptly characterize most of the attempts to naturalize intentionality. I don't think such a justification can be given.

However, I don't think such a justification *needs* to be given. To naturalize intentionality we can embrace the normativity at the foundations of intentionality, and rather than giving naturalistic *justifications* for particular norms, give a naturalistic *explanation* for how norms in general arise (whatever those norms happen to be). I'm going to tell a naturalistic *Just So Story* in Chapter Six about how (possibly arbitrary) norms came to be. The point for now, though, is that attributing specific contents to representations, contents that enable us to distinguish Type One from Type Two cases, appeal to norms about what the representation *should* represent.

Fodor's objection is that if a certain type of object *can* cause the representation to be activated, then this type of thing can rightly be said to be a member of the set of things it correctly represents. The problem is that we *want* to be able to come up with a justified, naturalistic description of the things the representation correctly represents, so that some things that can cause its activation are things it does not correctly represent. Since we're sure that representations *do* misrepresent, we think that we *should* be able to come up with

[39] Here I include functional analyses –analyses about what a particular device or item *should* do– in the realm of the normative.

such an account. As I will explain shortly, however, the only way to specify things that can cause the representation to be activated as things that the representation does not correctly represent, is to employ normative vocabulary: although they can cause the representation to be activated, they should not be included in the content of the representation. Such cases count as cases of misrepresentation, by virtue of a normative standard of what counts as correct and incorrect here.

2.4 Types of Representation.

Solving the disjunction problem could easily be the subject of a dissertation by itself. But we do not, as Fodor thinks we must, have to *solve* the disjunction problem. We simply need to *avoid* it. We just have to avoid calling cases to which the disjunction problem applies, cases of "representation". We will not find examples of misrepresentation unless we rule out cases where external events directly cause internal events and where these internal events directly cause behavioural responses. (It's the causal laws about the types of things that can cause the representation to be activated that give us trouble.) This is especially so for creatures that show no evidence of *cognitive penetration* of the perceptual process (i.e. where knowledge cannot alter the way the creature is disposed to respond). Only certain types of mechanism can count as representations that can be activated in the absence of the things they stand for. Haugeland's point (2) rules out internal states that cannot do the job of standing for external states of affairs when those states of affairs do not obtain. Points (2) and (3) also rule out internal states of creatures whose internal states are not part of a representational *system* of intentional states. This, combined with the previous remark, means that the perceptual states of beings that cannot act with regard to absent or temporally distant states of affairs, do not have content. They are caused by, but cannot misrepresent –and so do not represent– their objects. There is no *identification* of the object, but simply an automatic behavioural response to a certain type of perturbation of the creature's system, that this response is "tuned" to. Thus, frogs' perceptual states are just the kind of states that do *not* have representational content. The internal states of the frog that snaps at bee-bee pellets are similar in kind to the internal states of Brooks' robots; both of these are similar in kind to the Watt governor's operation. These types of system have feature-detectors whose triggering is caused by certain types of external perturbation, and which trigger an automatic behavioural response from the system. But because these systems operate purely in terms of a causal regularity, they cannot be caused to be activated in the absence of the kind of thing that causes them to be activated.

I'm not simply appealing to Haugeland's definition of a representation, however. The kind of definition Haugeland gives of a representation, apart from showing the restricted sense of representation that many anti-representationalists appeal to, also has support from a distinction made by Charles Sanders Peirce.[40] This distinction can help explain my recommendation that we avoid the disjunction problem and that intentionality is inherently normative.

Peirce makes similar remarks to Haugeland's, about the kinds of public signs he calls *symbols*. Symbols are a special kind of sign, for Peirce; as distinct from *icons* and *indices*. The difference in the types of sign, for Peirce, depends on the sign's *interpretant*. Very loosely, the interpretant of a sign is how the sign is used; the "mental effect", or the "cognition of a mind" (2.242) that it brings about in a person when the person interprets the sign. Here I'll only give a sketchy description of the difference, for Peirce, between these types of sign. *Icons* are signs that are interpreted to have a resemblance relation (2.282, 3.556) with the thing that signify (maps are used as icons of the areas they are maps of; a paint store's color swatches are icons of the colors of the paint they sell; today's firing of the cannon at dawn is an icon of yesterday's firing of the cannon at dawn). *Indices* are taken to have what Peirce calls an "existential" connection (2.243) with the object they signify; this is usually taken to be a causal relationship (a knock at the door is an index of the presence of a visitor; smoke is an index of fire). *Symbols* are related to their objects by virtue of a convention accepted by a community (2.246). Referential words of a language are good examples of symbols. Members of the linguistic community take each symbol to be a symbol of a certain thing, because of a (possibly rather arbitrary) convention of using that symbol to refer to such things. Thus to speakers of English the mark "dog" on a piece of paper is used as a symbol for the hairy canine pets some people

[40] Peirce (1960), esp 2.274-2.308. All references to Peirce will be to the volume and section numbers in this five-volume set. See also Von Eckardt (1993, section 4.3) for a good introduction to and explanation of Peirce's system of signs.

keep, and it signifies these creatures because of the accepted convention that this is how one is to use the word.

The significant point here is that symbols are the only type of sign that can maintain their relationship with what they signify in the absence of the thing signified. An icon cannot be taken to look like something that doesn't exist: "the statue of a centaur is not, it is true, a representamen if there is no such thing as a centaur" (5.73). Similarly, an index cannot be caused by something that wasn't there to cause it.

> An *index* is a representamen which fulfils the function of a representamen by virtue of a character *which it could not have if its object did not exist*, but which it will continue to have just the same, whether it be interpreted as a representamen or not. (5.73, my emphasis)

For example, smoke has an existential relationship with fire. The smoke would not be present were it not for the fire that causes the smoke. The smoke would be present and interpretable as an index of fire, however, even if nobody so interpreted it. However, with symbols, in contrast, members of a community can conventionally use them to stand for something that doesn't exist, or for something that is not related to the sign causally nor by resemblance. In this way an icon can be used symbolically; conventional ways of interpreting the icon are necessary to determine which features of the icon are taken to be the ones that are similar to the object. Thus the statue of the centaur can function symbolically, as a conventional sign of a general type of creature which, if it did exist would look like the statue.[41]

Peirce's distinction is about *public* signs. But the distinction can be used just as well to talk about mental signs (neurological "representations") and their interpretants. Doing so has the advantage of reinforcing Haugeland's narrower definition of a representation. Classical cognitive scientists and philosophers of mind maintain that all cognition is representational. Thus they assume that the

[41] The statue fails to be an icon because it does not represent the centaur by virtue of its resemblance relations alone, but by virtue of the convention that centaurs would look like a being with a horse's body and the upper torso of a human, if there were such a thing. A photograph of me, however would function iconically, since someone who viewed the photograph and also had actually seen me, would recognise the resemblance relation without the need for conventional support. That is, they would need no conventional support apart from the convention that the two-dimensionality is not one of the relevant ways that the photo is supposed to resemble its object (4.418).

perceptual states of frogs, bees, magnetosome bacteria, and so on, are representational states and attempt to give accounts of their content. However, according to Haugeland's definition, they are not representations. On Peirce's system, if they are representational at all they are representational in a quite different sense from the kind of symbolic representations that are supposed to account for human cognition. They may be indices, if the behaviour of the system uses it because it has a *causal* relation to the thing they signify, but they are not interpreted symbolically. These perceptual states have an "existential" relation with their objects, such that the state would not have the features it has were it not for the presence of the object. Thus defenses of the representational content of *perceptual* states of creatures that cannot act with respect to absent or temporally distant entities –such as that of the content of the frog's visual system when it sees a fly (or a bee-bee or a fleebee)– are defenses of an indexical system best described in terms of a causal regularity. They are indices of "something with the power to cause this state". On Peirce's system, they have an indexical relation with their signified objects that does not hold unless the signified objects are present and causally efficacious in producing that state. On Haugeland's account of what counts as representational, such perceptual states do not represent. Unless they are part of a symbolic representational system, it is an index, not a symbol, and cannot misrepresent. It is caused to be activated by the presence of a certain set of objects, and there is no principled way to describe this causal regularity so that some of the objects that cause the representation to be activated are things that it does not represent. The only justifiable way to specify this causal regularity is with the most general description: one that covers all the states of affairs in which the representation is activated. So the frog snaps at "small ambient moving black things" and does not misrepresent when it snaps at a bee-bee pellet.

In addition, the interpretant of such indexical representations is an action, not a concept or mental idea. The activation of the feature detector in the frog, for instance, *causes* the frog to respond behaviorally, by snapping. Peirce distinguishes such actions from the interpretants of symbolic representations (ideas, beliefs or concepts, that result in habit-changes), by virtue of their lack of generality. An action, he says "cannot be a logical interpretant, because it lacks generality" (5.491) One of the distinguishing features of symbols is their interpretant's generality. Symbols are general types, instantiated by *replicas*

(2.249), or tokens, that are either utterances or inscriptions of linguistic symbols, or the neurological instantiations of mental symbols. The objects of symbols are also "of a general nature" (2.249). For instance, the word "dog" and my concept of dog both refer to a *type* of object, rather than to particular instances of that type. The objects that are the objects of the indexical representations activated in frogs and magnetosome bacteria, in contrast, are not general types, but are *particular instances*. In Twardowski's (1894/1977) terminology, they have an object but no content. Each activation of a frog's indexical perceptual state has as its object the particular small ambient black thing that caused its activation. It is an accurate index of its object, and accurately guides the frog's snapping action towards that particular object (even if the frog misses, it was that object that the frog snapped at). But the frog does not represent that object, in the symbolic sense of representation which allows for misrepresentation.

2.5 *Norms and Symbolic Representations.*

The types of representation that do have content and can misrepresent, are symbolic[42] representations; representations that satisfy Haugeland's definition. To naturalize content, then, we need to find a naturalistic explanation for the content of this type of representation.

However, there are two potentially large problems with taking a Peirceian account of symbolic mental representation to be naturalizable. The first problem comes from that fact that the *interpretant* of the sign and its relation with what it is a sign of determines what kind of sign it is. For Peirce, the interpretant of a public sign is the "mental effect" it has on the interpreter: it calls up an idea or concept in the interpreter. This idea or concept is another symbolic representation. The regress problem here is obvious: we cannot call something a representation (of a certain type), because of its having a (certain type of) relation with the representation's interpretant, if that interpretant is itself a representation in need of interpretation. This potential chain of representations (even if of different types) needs to be grounded somewhere, in an interpretant that is not itself a representation and in need of interpretation.

[42] Note this is "symbolic" in Peirce's sense. Vera and Simon use "symbols" to stand for any type of representation, iconic, indexical or symbolic. Because of this potential confusion, I'll stick to using "representation" to refer to the general class, and "symbol" (in Peirce's sense) to refer to the subclass.

Peirce's solution to the regress problem is to point out that interpreting a sign can cause a mental effect, which can cause a mental effect, and so on, but not all the subsequent effects will be mental ones. In "A Survey of Pragmaticism" (5.476 ff.) Peirce argues that the "logical interpretant" of a thought or "mental sign" cannot be another mental sign (due to infinite regress) (5.476). The only mental effect that is not itself a sign (and so doesn't have an interpretant) is a *habit change* "...meaning a change in the person's tendencies toward action resulting from previous experiences or from previous exertions of his will or acts, or from a complexus of both kinds of cause". Peirce takes this to be a ground for the regress, in that the chain of representations that are the interpretants of other representations comes to an end at something that is not a representation and is not in need of an interpretant. People's habits of action, for example, are also the result of their interpreting a sign and coming to hold a new belief. Having this new belief will alter the person's dispositions to act, making them disposed to do or say certain things in certain situations. The actions the person is now disposed to perform are at the end of the causal chain. They ground it in something that is not itself a sign in the way a mental representation is a sign. Although an action can be interpreted, says Peirce, it is not *in need* of interpretation in the same way that the mental representation was (5.491). Peirce's solution to this regress problem, by relating all mental representation to the representer's actions is an important part of the embodied action picture of naturalized intentionality I'm developing; one that leaves a lot to be said however. I'll come back to this presently. But first I need to talk about a further problem with the attempt to give naturalized accounts of symbolic representation's content.

The second problem with naturalizing the intentionality of a symbolic mental representation is more serious. (It also undermines, as I'll argue in the next chapter, Peirce's solution to the regress problem). This problem arises because the referential relations of symbolic representations, to Peirce, are supposed to depend upon a *convention* accepted by a community.

Symbols are the only kind of sign that can misrepresent; they are the only kind that do not have an "existential" relationship with their objects. Symbols are the only kind of representation that have contents as well as objects. Icons and indices only have objects. This is because symbols have their content –their meaning or their intentionality– by virtue of conventions; there are norms of how the sign ought to be interpreted. These norms governing the relationships among symbols are what enable symbols to stand arbitrarily for things that do not exist or that do not have causal or resemblance relationships with their objects. Symbolic systems are conventional through and through.

Barbara Von Eckardt (1993) makes this point explicitly, saying "the existence of a conventional ground [for a neurological symbol's content] is ruled out at the outset because of cognitive science's commitment to naturalism" (p. 206). Conventions are not natural, for Von Eckardt. They depend on the intentional states of agents. Additionally, giving a naturalistic *justification* for a convention (i.e. a norm) is supposed to be impossible. It is subject to Hume's (1739) objection that one cannot derive an "ought" from an "is"; naturalistic explanations of how things are cannot support normative claims about how things ought to be. This is why, as Von Eckardt goes on to point out, explaining mental representations in terms of an indexical or iconic ground "is at the heart of most of the current approaches to the content-determination question " (p. 206). She also points out (p. 410, note 6) that Fodor (1984, p. 233) makes the even stronger claim that theories based on causality (i.e. indexical theories) and theories based on resemblance (i.e. indexical theories) are the only two naturalistic theories of representational content that have ever been proposed. This perceived need for a non-conventional ground for the content of a mental representation is the reason Von Eckardt endorses Peirce's solution to the regress problem. The ground, for Peirce is a habit change; a disposition to perform certain actions. These actions are not in need of symbolic interpretation.

A further problem for conventional grounds for the representations that most philosophers of mind and cognitive scientists agree are the basis of human cognition, is that these representations are supposedly inside people's heads. Such bits of people's neurological apparatus are hardly accessible to public scrutiny, and cannot feature as part of shared normative conventions for what they are supposed to represent.

This restriction on naturalistic explanations –that conventions are disallowed as explanations in naturalistic theories, since conventions are allegedly not naturalizable, because they themselves depend on the intentional states of agents– is also the reason, in my opinion, that nobody so far has given a satisfactory naturalistic account of representational content.

I disagree with this restriction. The point of the next chapter is to argue that conventions and normativity are at the heart of any theory of intentionality, but also to argue that this is not fatal to a naturalized account of intentionality. I'll give a brief explanation for this claim now, and come back to explain more fully towards the end of the next chapter.

The intentionality of our public symbols is often supposed to be derived from the intentionality of our mental representations. Conventions about the concept that one should associate with a public symbol are harmless enough in themselves. But when this combines with the common predilection to see intentionality as attaching to neurological representations that instantiate our concepts and intentional states we run into trouble. Conventions cannot govern hidden internal states of people's neurological apparatus (not until brain-imaging apparatus is in much more common use, at least; see Section 5.2 for discussion of the problems, even then). And furthermore, conventions are supposed to *depend* on the intentional states of the individuals who make conventional associations and who enforce such conventional associations. Thus intentionality cannot be explained in terms of conventions, if conventions are themselves intentional in nature. In spite of these common misgivings about an account of intentionality based in conventions and norms, I am going to argue that the intentionality of people's thoughts, intentions, beliefs and desires is deeply conventional in origin. I am going to argue that mental intentionality, like linguistic intentionality, is derived from the intentionality instituted by the public norms implicitly and explicitly governing the shared practices that comprise human beings' forms of life.

Dennett can give us a start on seeing the extent to which mental intentionality is based in normative practices. Dennett (1971, 1987, 1991a) argues that all intentionality is in the eye of the beholder. Whether a creature or system has intentionality depends on whether an observer of that system adopts the intentional stance towards it, whether the observer attributes intentionality (i.e. purposes, goals, beliefs, desires etc.) to it. Ultimately pragmatic success justifies the observer in attributing intentionality to the system. Adopting the intentional stance towards a system, and explaining its behaviour in terms of its intentional states is justified if this stance enables one to make more successful predictions of the system's behaviour.[43]

An extension of Peirce's solution to the regress of representations problem works in the opposite direction, coming to the same conclusion. Peirce terminates the regress in something that was not itself in need of interpretation: Habit-changes. Habit changes are changes in the agent's dispositions to behave, as evidenced in the agent's actions. Although Peirce thinks that these actions are not in need of interpretation, these actions *are* interpretable. In fact, this interpretation of others' actions in order to predict their behaviour, is based on taking their actions as *signs* of the intentional states of the agent. Such interpretation is the basis of Dennett's entire position on intentionality. An agent's actions are the basis upon which others attribute beliefs, desires and intentions to the agent. Successfully predicted actions of an agent *justify* attributing such beliefs, intentions and desires to the agent. Further actions of the agent are also interpretable as signs of those internal states, and can confirm those attributions of intentionality.

These attributions and confirmations work, however, not simply as Dennett supposes because it is pragmatically useful. The pragmatic benefit of this is a side-effect of the fact that the connections between actions and attributions of intentional states are normatively enforced. Think back to the distinction between indexical internal states and symbolic ones. Indexical signs have a causal relationship with what they signify. Symbolic signs are conventionally related to what they signify. Attributions of intentional states help make predictions about what an agent can be expected to do. These are not based on causal regularities governing what the agent in that intentional state *will* do. Rather they are based, as Brandom (1994, p. 56) points out, on norms that license conclusions about what intentional state we *should* attribute to someone based on particular actions, and norms about what an agent in that intentional state *should* do. Our folk psychology is a normative theory: it licenses attributions of intentional states as *reasons* for people's actions, not as *causes* of

[43] For example, Dennett says: "...there could be two different, but equally real patterns discernible in the noisy world... [e.g. one observed from the intentional stance one from the physical stance]. The choice of a pattern would indeed be up to the observer, a matter to be decided on idiosyncratic pragmatic grounds" (1991a, p. 49).

their actions.[44] If the agent says that he wants the light on, then we should attribute to that agent the desire that the light be on. If the agent does have this desire then the agent *should* do something that will bring it about that the light is turned on (thus their desire also serves in some cases as a reason for their declaring this to you).

> To say this is not yet to say that the one who has such a reason *will* act according to it, even in the absence of competing reasons for incompatible courses of action. What follows immediately from the attribution of intentional states that amount to a reason for action, is just that (cetirus paribus) the individual who has that reason *ought* to act in a certain way. This 'ought' is a *rational* ought—someone with those beliefs and desires is rationally obliged or committed to act in a certain way. (Brandom 1994, p. 56)

The norms of the practice of giving and asking for reasons are supports for our practice of, as Brandom puts it, keeping "score" on one another and what we expect each other to be committed to doing and saying. If you know that Mason wants to finish his dissertation, and that he believes that he needs to spend every waking moment in the next few weeks working on it in order to finish it, then you are licensed by the norms of the practice of attributing these desired and beliefs, to have expectations about what Mason ought to be committed to doing for the next few weeks. Thus Mason's going away for the weekend to relax in the sun would be going against that commitment.

One of Brandom's (1994) central points is that attributing intentionally contentful states has normative consequences; "intentional states and acts have contents in virtue of which they are liable to evaluations" that form the core of "the social practices of giving and asking for reasons" (p. 17). The shared social practice of giving and asking for reasons for actions, institutes norms of inference (the *propriety* of certain inferences).

These norms license inferences from the actions agents perform to particular intentional states one is justified in attributing to the agent as the

[44] Davidson (1963) argues to the contrary. The agent's reasons are causes of actions to Davidson; "rationalization is a species of causal explanation", he says (p. 3). Davidson uses this to motivate a token identity theory, in which reasons are events that under a different description are the physical events that stand in causal elation with the action. Here Davidson is explicitly disagreeing with "Wittgensteinian" accounts (see p. 10) like mine. Explaining why Davidson is wrong about this is outside the scope of this work, however. On this, we will have to agree to disagree, and leave it at that for now.

agent's reasons for performing that action. Saying "I wish that the lights were on" is interpretable, by virtue of our linguistic conventions, as a symbol of the intentional state of desiring that the light be on. Similarly, getting up out of one's chair and flipping the light switch is also interpretable as a symbol of this intentional state to that person. The criteria, then, for being in a certain intentional state are public, shared criteria.

The norms that license attributions of intentional states also rule on the propriety of inferring what actions someone ought to perform if such and so intentional states are correctly (according to these norms) attributed to them. If one has certain beliefs and desires, then it would be rational to act in such and so a way; they *should* act in that way. If I want the light on and believe that flipping the switch will turn the light on, then it is rational for me to turn the light on. This is not to say that I *will* turn the light on, however, but that I *should* turn the light on. The norms for the uses of the intentional expressions we employ to do so, stipulate certain behavioural criteria that must apply when they are used. If I have these intentional states then I *ought* to act that way. And, conversely, if I act that way, my actions can be interpreted as symbols for my having those intentional states that would be reasons for my action.

Thus people's actions are not, as Peirce supposes, a non-representational ground for mental representations. Rather actions are interpreted as *symbols* of people's intentional states. They are not indices, caused by the agents intentional states, but symbols, whose relationship with the intentional states they symbolize is based on the norms and conventions of the practice of giving and asking for reasons for actions. The regress of symbolic representations, then, does not get terminated at something that is not a symbol. The actions that Peirce presents as the termination of the chain of symbols interpreting symbols, interpreting symbols... are themselves symbols, and are interpreted as such. Human lives, and the mechanisms that explain them are symbolic, and thus normative, through and through. I'll have much more to say about the normative nature of attributions of intentional states in the next few chapters. But for now, lets think about the relationships between these practices and the intentional states that are ascribed to people, and the symbolic representations posited in cognitive science explanations.

2.6 *Ascriptions of Intentional States and Representations*

As I have been explaining, I disagree with the requirement that naturalizing intentionality entails naturalizing the content of neurological structures that are people's representations. As I've been explaining, if there are representations that explain people's cognitive abilities, they must have contents, and thus be capable of misrepresenting. But the only kind of representation that has content and can misrepresent is a symbol. Indices and icons have "existential" relationships with their objects. The problem here is that symbols are related to the objects by virtue of the practices within which the symbols have their life, and the norms and conventions of interpretation those practices institute. And although I do think that conventions and norms can be given a naturalizable explanation (this is the point of this dissertation, as we'll see) the content of such representations is not possible to naturalize in this way. This is because these neurological items *themselves* do not play any role at all in the kinds of practices that confer contents. The representations posited in cognitive science explanations are hidden, internal states of people's inner neurological workings.

The Embodied Action approach to cognitive science rejects some of the foundations of this way of phrasing the problem; a way of phrasing it that makes internal mechanisms which *are* contentful representations be the solution. It also suggests an alternative; one that may be more soluble. Let's take the problem from the beginning, to see the basis of this rejection and the basis of the alternative. We are agreed that the aim of cognitive science is to explain how it is that human beings' are able to do what they do. One of the capacities human beings –perhaps only human beings– possess, is the ability to deal with what Clark calls representation-hungry problem cases; we can think about and act with respect to things other than those that are presented perceptually in our current environments. One traditional step in the explanation of these capacities is to posit mental states; desires, intentions and beliefs of agents. The next step is to ask what these intentional states really are. The answer is given that they must be states of the brain. One of the central points of this dissertation is to focus attention back on these capacities to solve representation-hungry problems, and away from the representations that are posited to explain these abilities. I want to focus our attention back on the capacities to perform actions that are directed at counterfactual or absent or abstract states of affairs. People desire to have a dinner this evening that the kids will enjoy, plan what to cook for dinner tonight, remember that the kids like macaroni and cheese, worry that there is no cheese left in the house, try to remember to stop at the store on the way home to buy cheese, and try to get home in time to make dinner at a reasonable hour. The problem comes when we move from such things people are able to do, to the intentional states that we suppose must lie behind them, carrying the assumption that such intentional states are really states of the brain. This latter assumption gets the problem of intentionality characterized as a problem about how one piece of the world (that is, a piece of a brain) can be directed at or about another piece of the world.

The question, however, could just as justifiably stop at these actions themselves and the intentional states that are the reasons for performing those actions. I worry that there is no cheese in the house, because I believe that I used the last of it to make a sandwich for my lunch today. I try to remember to stop at the store, because I want to have dinner ready and I know that I won't if I forget to stop to buy cheese.

What is it that makes these actions and intentional states directed at certain states of affairs? What explains the agents reasons for doing as they do? What makes these intentional states have the contents that they have? I'm going to argue that the answer to this is *norms*. The norms of human practices make it the case that certain actions people perform are directed at particular objects and states of affairs. Or rather, these norms make it that case that these actions *count as* being directed at certain states of affairs.

This is where we'll find intentionality "at home". It is not to be found in people's internal neurological states and mechanisms. Intentionality is a property of people's actions and of the intentional states that are attributed as the reasons for those actions. These actions and intentional states have roles to play in people's norm-governed social practices. The norms confer intentionality on these actions and intentional states, by virtue of these roles they play in the practices.

Now if it turns out that these internal neurological mechanisms come to have roles to play in people's norm-governed social lives, then those neurological items may come to have intentionality conferred on them also. For instance, imagine that my ability to recall whether there is cheese in the refrigerator is due to a neurological state which functions to enable me to recall what is in the fridge. People with lesions to this bit of their brain are unable to recall what is in their fridges. Whenever I think about what is in my fridge, brain scans show a high amount of activity in this area of my brain. Eventually even perhaps the causal mechanisms underlying the ability to recall what is in the fridge can be traced to properties of this area. If you show me that, contrary to my expressed belief, there *is* cheese in the fridge then the properties of this area change in what come to be predictable ways, and so on. This mechanism enables me to think about the contents of my fridge.

I am going to argue that this item is not *itself* my belief about what is in my fridge, although it may be causally responsible for my ability to perform actions that license attributions of such a belief. This item enables me to perform actions that count as directed at the contents of my fridge: to worry that there's no cheese in the fridge, to hope that nobody has drunk that last beer I recall noticing in the fridge at breakfast this morning, and so on. By virtue of its role in enabling such actions we may want to say that it derivatively has intentionality. But the true possessors of intentionality —the intentionality conferred by the normative practices of describing people's actions, and of giving and asking for reasons— are the actions this item makes possible, and the contents-of-the-fridge-directed intentional states that count as the reasons for such actions. These intentional states are predicated of whole embodied persons and their actions. This intentionality is derived from the norms of the practices in which such actions and states have their "life". Whatever intentionality we might confer on internal mechanisms –once they enter the sphere of people's norm-governed lives and the explanations for people's actions– will be derived to a second degree; derived from the actions that the mechanism makes possible.

An analogy with linguistic intentionality might help to make this clear. (I'll be talking a lot more about this in the next chapter.) According to the kind of embodied action view I'm outlining here, people perform speech acts that count as being directed at certain objects and states of affairs. Words are tools that have conventional uses, and without which such speech acts would not be possible (or at least not as easy). Without phases such as "in my fridge" I couldn't perform the speech act of informing you about what is in my fridge. But the phrase "in my fridge" is not *itself* directed at the contents of my fridge. It's an item that gets its life within certain speech acts, and it enables those speech acts to *count as* being directed at the contents of fridges. But the intentionality here is the instituted intentionality of the speech act; intentionality conferred by the practice of interpreting such speech acts as being directed at the contents of fridges. Whatever intentionality we might confer on phrases such as this, will be derived to a second degree; derived from the derivative intentionality of the speech acts within which it has its "life".

On this view, then, there is no such thing as "intrinsic" intentionality, and no objective (practice independent) way to specify what makes something have intentionality. All intentionality is derived; derived from the practices and norms that permeate and support human forms of life.

The way to naturalize the intentionality of human intentional states and actions, then, is not to reduce it to non-intentional explanations, nor to terminate or ground the intentionality in non-intentional physical phenomena. The way to naturalize intentionality is to give a naturalistic explanation for the human social practices in which the norms that confer intentionality on people's actions, and on the items and structures that play a role in these practices, have their origin.

2.7 *What is a "naturalistic" theory?*

Before I proceed with giving such an explanation in the next chapter, it is important to pause to give a reminder about what it is to be a *naturalistic* explanation for a phenomenon. The point of a naturalistic explanation of a phenomenon is simply that no appeals to magic (exemptions from the laws of nature) are involved in the explanation. A naturalistic explanation of a set of phenomena brings the study of those phenomena into the study of the world of nature. It explains those phenomena by using the methods and tools employed in the study of the natural world —that is, they must be given a scientific explanation. In such explanations, no entities apparently outside of the natural realm can feature. If such entities do feature in an explanation, they also must be explained away, by integrating them into a science: by explaining them while appealing only to entities that can be given a natural, scientific, explanation. Attempts to explain away such phenomena, cannot, of course, appeal to the phenomena we are trying to explain away. Such circularity is the bugbear of most contentious naturalistic explanations. The charge against most naturalistic explanations of intentionality, for instance, is that they attempt to explain intentionality, while appealing to intentional phenomena.

Explanations of human capacities that invoked Cartesian minds, did appeal to such "magic"; they appealed to entities that explicitly were exempt from the laws of physics. Physicalist philosophy of mind has been trying to find a way to explain humans' capacity to think about absent objects, to plan future contingencies, and so on, without resorting to appeal to magical entities such as Cartesian minds. The way many theorists attempt to naturalize intentionality, and the way most argue that it must be naturalized, is to assume that the mind is really a physical system –the brain– and thus to *reduce* explanations about the operations of minds invoking intentional concepts to explanations invoking only the concepts and explanations involved in explaining the behaviour of physical systems; i.e. physics, chemistry, biology.

The way to "naturalize" intentionality, however, is not to explain intentionality away, to *reduce* it to physics. As Jeff Foss (1995) explains explanatory reduction of one field to physics is not necessary for that field to have scientific respectability. Indeed, this rarely happens. What is necessary is not this extremely rare kind of *explanatory unity* with physics (physics making the other theory redundant), but *ontological unity* (the unity of entities posited in theories, such that no theory invokes entities that physics rejects as "magical"). Foss explains ontological unity in terms of the *information economy* of science. Different sciences trade information and methodological techniques, experimental equipment and techniques of information extraction with one another. Ontological unity, says Foss (p. 419), "flows from the universal applicability of physics: no matter how arcane the special theory or discipline, there is always the possible, and usually the actual, receipt of information from physics." Foss mentions DNA theory, and paleo-anthropology as examples of sciences that trade information with physics. Evolutionary theory is another good example of a science that trades information with physics, and with other sciences that trade information with physics. Evolutionary theory and the study of the way species and their environments evolve through natural selection, is informed by fossil records, carbon-dating, biological studies of such records and of the present biological mechanisms and their probable evolutionary history, the mechanisms of genetics and many other scientifically respectable fields and their explanatory theories and experimental techniques. Cognitive ethology is another such scientifically respectable field, that explains the behaviour of animals and the mechanisms that account for animals' abilities, while trading experimental techniques and theories with other sciences, including biology, chemistry and physics. If explanations in fields such as this can account for the entities and the phenomena we employ in explanations of intentional phenomena (i.e. practices and norms), then we can claim to have "naturalized" intentionality.

Furthermore, if these fields could give a scientific, causal explanation for how such norms came to be, then these norms would be definitely shown to be naturalizable.

They can give such explanations. We need to appreciate that norms of behaviour are respectable entities, that can and do feature in scientific explanations of the behaviour of creatures. Cognitive ethology and evolutionary theory both traffic in norms of behaviour. Cognitive ethology employs norms to explain the way animals are conditioned into conforming their behaviour with the practices of their groups (their herd, flock, hive, troop, etc.). Evolutionary theory uses norms as one way of explaining how natural selection can select between groups with different ways of behaving, when those differences produce varying "fitnesses" of groups. Norms are a feature of the scientifically respectable explanations found within these fields; fields that participate in an information economy with physics. And these fields also explain the way the norms arose; they give an evolutionary Just So Story to explain how norms were caused to be the way they are. I'll say much more about this (among other things) in Chapter Six. For now I'll give a brief introductory defense of this claim.

The behaviour of members of flocks of birds, herds of cattle, and tribes of apes is often subject to normative restriction, as Haugeland (1982, 1990) points out. The process of bringing the behaviour of young members of a group into conformity with the rest of the group, through behavioral conditioning by conspecifics, explains much of the behavior of such animals; they *learn* to behave that way, rather than being genetically programmed to do so. They *could* behave differently, but after a period of socialization where certain behaviours are reinforced and others are discouraged, they don't behave differently; they conform to the ways a member of this flock, herd, tribe, etc. ought to act. The "cultural" practices of different groups, and natural selection between separate groups with different practices that enable the group to prosper to varied degrees, are often cited as a source of group selection (e.g. Sober 1984, 1991).

The solution to the problem of a naturalizable ("scientifically respectable") explanation of intentionality, is not to reject the normative nature of intentionality as "non-naturalistic" and in need of reduction to non-normative explanations. Rather the best tactic is to explain intentionality by appealing to the norm-governed practices in which such intentionality is instituted, and give an evolutionary explanation for how such practices arise. As I said earlier, my point is that the existence of norms can be *explained* naturalistically, even if the norm itself cannot be *justified* naturalistically. Various practices, and the norms they institute need not be justified. Just as evolution by natural selection does not imply the intentionality of a designer who made creatures with certain traits, so the cultural evolution of practices does not necessarily involve the intentions of agents who design practices. Most norms and practices were simply "stumbled upon". Errors occur, where behaviour that was once censored is allowed to let pass, and other behaviour that was let pass begins to be censured. These "errors" that sometimes are reinforced and lead to alterations in the group's practices that may be selectively advantageous to the group, correspond to the "copying errors" that lead to alterations in genetically produced traits that are selectively advantageous to an individual. By these means the fact that such-and-such a norm prevailed can be explained naturalistically, even if it cannot be so justified, by appeal to the selective advantages of a group whose members conformed to the practice that institutes that norm.

Of course even though many norms were simply "stumbled upon" back in pre-prehistoric times, sometime in our evolutionary history our ancestors became able to consciously design and improve our norms and practices. Our practices now have, at least sometimes, conscious intentional design behind them. A trivial example is the practice of designing and making better tools, and teaching young folk how to make them too. This ability to think about the practices themselves, and the ability to improve them through conscious design, is an important evolutionary step.

It is especially important to give a naturalistic explanation for the origin of the practice of attributing intentionality to others. As Dennett points out, intentional systems are such only because something takes the intentional stance towards them. There would be no intentionality if there were not beings adopting the intentional stance. Less individualistically put, all intentionality is instituted by a community of intentional systems' practices of treating things (including one another) as though they have intentional states.

Explaining this ability to adopt the intentional stance, then, is the crux of an explanation of how intentionality is possible. Explaining how human beings became the kinds of creatures that can *explicitly think about* the content of our norms, can break them intentionally, and reform them, is going to be necessary if we are going to give an evolutionary explanation for the kinds of norms that give rise to intentionality. This ability, I will argue, is based in human beings' ability to use language. I will argue that the ability to use language, and the ability to think about and attribute purposes and intentions to oneself and others (the ability that gave rise to intentional states as explanations for why certain creatures do what they do), evolved together. The co-evolution of human beings' folk psychology, language, cultural practices and cognitive equipment gave rise to human beings' present range of abilities.

In giving a naturalistic explanation for human beings' abilities, then, two steps are crucial. First, to give an evolutionary explanation for how our ancestors came to be able to adopt the intentional stance; for the conditions under which this ability was selectively advantageous for individuals within a group, and for groups as a whole. Second, we need to explain how this ability enabled the development and flourishing of linguistic interactions that are based on the shared practice of attributing intentional states to others as reasons for their (speech and other) actions. I am going to argue that the simultaneous evolution of these capacities –and their co-evolution along with the brain structures and the social and linguistic practices that enable us to exercise these capacities– "bootstrapped" human beings and our cultures to bring about the present situation, and our present cognitive capacities. Giving an account of how this happened, and how it might naturalistically explain the capacities that we appeal to intentionality to explain, is the aim of this dissertation.

The next chapter begins this task by giving an account of the intentionality of language and how linguistic practices and their norms shape the attribution of intentional states to people.

Chapter Three

Intentionality in Practice

But as soon as he began thinking what he was doing and trying to do better, he was at once conscious how hard the task was, and would mow badly.

–Leo Tolstoy (1960)

In Chapter One, I introduced embodied approaches to cognition, where the focus is on agents and what they can do, rather than on minds and what they can think about. In Chapter Two, I discussed the disputed role of representations in cognitive science, and outlined the problem of naturalizing intentionality. I introduced the idea that this problem should not be seen as the problem of explaining how one bit of the world can be about another bit of the world. This is because intentionality does not attach to representations *themselves*. Rather people perform actions, and intentional states are attributed to people (including oneself) as reasons for the actions they perform. Thus intentionality —contentful intentional states— are predicated of people as they perform actions that are situated within shared social practices. Intentionality is not predicated, in anything other than a highly derived and abstracted fashion, of neurological mechanisms and states that enable people to perform the actions they perform.

In this chapter I will extend this idea, and go a little deeper into the thesis that the intentionality typically attributed to mental states is not –as it is typically supposed to be in physicalist philosophy of mind and cognitive science– at the sub-personal level of neurological items and their properties. Nor is the intentionality at the personal level of people and their intentional states. Rather, it is at the inter-personal level, of people's norm-governed intentional actions and interactions, social practices, conventions and customs. All intentionality, linguistic and mental, is conferred on people's actions and on people's intentional states that count as reasons for those actions by the norm-governed practices that we participate in together.

The point of this chapter is that such practices are inherently linguistic. It's through having language –linguistic practices and abilities– that we can make our intentional states explicit. The norms on the practice of attributing intentional states to others and to oneself are based in the linguistic practice of performing the speech act of ascribing such intentional states, and the felicity conditions on such acts of ascription. These conditions and the fact that people can indeed judge whether intentional states are felicitously ascribed, I will argue, make it that case that the correctness –or better felicity or propriety– of ascribing an intentional state to someone depends not upon the ascribee's internal states, but on their publicly observable behaviour.

In order to argue this point about intentional states and their ascription, however, I first need to present a little more detail about how speech acts fit into our social norm-governed practices, and thus how the intentionality of language arises only by virtue of these norms and practices.

3.1 *The intentionality of language and of mental states.*

It will be helpful to begin with a sketch of the range of basic positions on intentionality, and a common view of how utterances have intentionality. A good guide to the different ways theorists approach the intentionality of language and its relation to that of people's cognitive states, is Haugeland's (1990) taxonomy based on the fielding team in a game of baseball. He divides theorists into three principal approaches, at first- second- and third-base, with slightly "deeper" versions of each approach playing outfield of each base. Having a picture of this taxonomy of approaches will make it easier, later on, to illustrate the ways in which my position differs from various other theorists.

At first base, Haugeland situates theorists such as Fodor, Pylyshyn, Field, Block, Cummins, Harman, and Lycan (1990, endnote 10). These "neo-Cartesian" approaches (p. 388) maintain variations on the theme that intentionality is possessed by individual items of people's cognitive apparatus. As I showed in the first part of Chapter Two, a representation in this framework is a particular neurological item. The intentionality of the neurological item depends on the role that item plays in an overall system (a language-like system, for some) of such intentional states. Haugeland situates Searle "outfield" of this position (p. 387, see also endnote 9).

Second base approaches, which Haugeland dubs "neo-behaviorist", or "mid-field phenomenologist" (p. 395), differ from first base approaches in that they ascribe intentionality to the overall cognitive system, rather than to its internal components. Thus we ascribe beliefs and desires to whole people without there necessarily being items in people's neurological apparatus that are those beliefs and desires. Such attributions of intentional states are based on the "environmental interactions" of the system, and whether those interactions can be properly characterized as the "competent" actions of a rational agent (p. 398). Here we find theorists such as Quine, Dennett, Stalnaker, Austin,[45] Bennett, Grice, and perhaps Ryle (p. 395, note 16; Haugeland can't decide whether Ryle "plays second base way back or centre field close in").

First base and second base approaches deal with the intentionality of language in similar ways. For both first base and second base approaches, says Haugeland, the intentionality of bits of language (words, sentences, etc.) is derived from a relationship between the words uttered and the mental or cognitive states of the speaker (p. 402). The fact that the speaker's mental states are directed at the world in the appropriate way –that the speaker's mental states are genuinely about the items in the world– anchors the intentionality of the piece of language associated with that mental state. (These mental states, of course, are either items in the agent's neurological apparatus or overall states of the agent.)

Some aspects of John Searle's account of the intentionality of language is a good example of the view shared by first base and second base theorists.[46] According to Searle (1994, pp. 78-82) the intentionality that objects such as pictures, words, sentences, maps, etc. are said to have, is *derived* intentionality. These things get whatever intentionality they have from the *intrinsic* intentionality of the objects' creators and interpreters. A sentence uttered by a speaker is only about some object in a sense of "aboutness" derived from the cognitive states –for example the speaker's intentions, to Grice– that the speaker associates with the utterance, which are *intrinsically* about that object. (For Searle (e.g. 1969, 1987), it is also a function of what the sentence uttered means in the language spoken. But this is a contentious point; not all first base and second base theorists would agree with Searle on this point.)

Searle also makes a sharp distinction between intrinsic intentionality and what he calls *as-if* intentionality. This distinction assumes a natural division between things that do and do not have minds. Human beings are said to have

[45] Perhaps because I base my view of language on Austin, I have reservations about situating Austin at second-base. It seems to me that although many aspects of his approach fit with second-base theorists (particularly his focus on agents performing actions), there are also reasons to situate him at third base.

[46] He is out in right-field because of his idiosyncratic views on mental intentionality.

"genuine" minds (and thus intrinsic intentionality), whereas if we attribute intentional states to things that do not have minds, this type of intentionality must be purely metaphorical. Searle uses the term *as-if* intentionality to refer to cases where we explain the behavior of objects (such as toaster-ovens, robots, ants and bacteria) by talking about them as if they had intentional states (beliefs, desires and intentions), even though they do not *really* have any such intentional states. This distinction does not rely on any proof of which things have and do not have intentionality. Rather it rests on intuitions that of course humans have genuine intentionality and toaster-ovens do not. Searle argues that such genuine intentional states are biological phenomena. Intentional states are emergent properties of brains and things with the same causal powers as brains; unhelpfully, these "causal powers" are "the brain's causal capacity to produce intentionality" (Searle 1980, p. 424). Thus utterances that are not caused either by the operations of brains, or by things with the same causal powers as brains, cannot have any intentionality supporting them.

Searle's depiction of as-if intentionality is principally an objection to (a caricature[47] of) Dennett's (1987) "intentional stance". Underlying the dispute between these two authors, is a basic disagreement about whether or not there is a fact about the matter, as to whether something *really* has intentionality or not. Searle is quite convinced that there is such a fact: an ascription of intentionality must be either true or false, and what makes it so is the presence or absence of genuine intentionality in the object it's attributed to (pp. 78, 82). To Searle, at least human beings have intrinsic intentionality. Dennett, in contrast, maintains that the object's behavior makes it sensible to *attribute* intentional states to the object, while denying that there is a fact about the matter about whether the object *really* has intentionality (in denying this he , along with Quine, are atypical of second-base theorists). Searle objects that this view faces the reductio ad absurdum that this means that everything in the universe has intentionality. Denying, as Dennett does, the difference between intrinsic and as-if intentionality, says Searle, "is absurdity, because it makes everything in the universe mental" (p. 81). Dennett, however, takes the other side of the denial of this dichotomy: nothing in the universe *intrinsically* has intentionality. All intentionality is *derived*, by being attributed by beings that adopt the intentional stance.

[47] See Searle (1994b), p. 81.

In Searle's infamous Chinese Room thought-experiment[48], he argues that if a system issues noises interpretable as spoken utterances or makes marks that are interpretable as written utterances, it is still an open question whether those utterances really have (derived) intentionality. Let's put to the side the interpreter's understanding of the utterance, and the way the mental states generated in the interpreter could give the noises or text intentionality.[49] For Searle, in order for these utterances to have the kind of derived intentionality that a person's utterances have, two conditions must hold. First, the words uttered must have certain meanings in the language spoken, according to the conventions of that language. And furthermore, the entity making the utterances must have the requisite mental states with intrinsic intentionality.[50] Otherwise there is no intentionality for the utterance's intentionality to be derived from. To Searle, for a system's utterance of "I would have you move from between me and the Sun" to be a genuine request, rather than just noise, two conditions must hold. There must be a *genuine* desire-that-you-move-from-between-me-and-the-sun present in the system causing the noises or marks to appear, and the words used must mean, according to the conventions of English, that the system has just that desire. An important criterion for utterances having derived intentionality, then, is their association with genuine mental states that have intrinsic intentionality in the agent making the utterance.

Davidson (1986) (who Haugeland (p. 418) rates as straddling first and second bases, but who might incorporate a touch of third base as well) gives us

[48] For example, Searle (1990)

[49] Based on the way Searle presents his account of meaning in his (1987) "Indeterminacy, Empiricism, and the First Person" we can do this. Here Searle argues that meaning something by an utterance is something that a speaker can do alone, independently of any interpreter's ability to determine that the speaker means this. What the speaker means is a function of what the sentence uttered means in the language the speaker is using, and of the speaker's intentions. The presence of interpreters who are capable of recognizing that the sentence (and thus the utterance) means this appears to be irrelevant to the speaker's meaning it, for Searle.

[50] See Searle (1969), p. 45 ff. The combination of these two conditions is exemplified in Searle's expressing the felicitous performance of a speech act as someone's "Saying something and meaning it" (p. 46); that is, meaning what the sentence means in the language. When he introduces his Principle of Expressability –whatever can be meant can be said (p. 20)– he says that "to study speech acts of promising or apologizing, we need only study sentences whose literal and correct utterance would constitute making a promise or issuing an apology" (p. 21).

another good example of this kind of analysis,[51] in his example of Diogenes saying to Alexander the Great the Greek equivalent of, "I would have you stand from between me and the Sun." Davidson says that Diogenes utters this, "with the intention of uttering words *that will be interpreted by Alexander as true* if and only if Diogenes would have him stand from between Diogenes and the Sun" (435, my emphasis). Interpreting Diogenes amounts to interpreting the intended meaning of Diogenes' words.[52] Thus Diogenes' intentions give the statement its meaning. Correctly interpreting Diogenes' utterance, amounts to interpreting it as he intended it to be interpreted. To do this, Alexander has to recognize what conditions would have to obtain in order for these words to express a true proposition. In this case, the truth condition of Diogenes' statement is Diogenes' mental state: this sentence would be true if Diogenes does in fact want Alexander to move.

Grice is also someone who fits this picture. For Grice, the intentionality of language is all based on the intentions of the speaker (or the intentions a speaker would have if they were to utter the sentence), and on the audience recognizing those intentions. The account I presented back in Chapter Two was based on Grice's idea.

One minor difference between Grice and myself is my focus on the action of performing a speech act, rather than principally on the utterance that I make. The second distinction between us is more important for the moment, and applies generally to all second-base theorists. (This is one of the principal reasons why I don't characterize myself as a second-base theorist. It's also the reason that I would deny that Austin fits comfortably at second-base.) The difference is the *absolutely central* role that norms and conventions play in my account. Grice (1975) does acknowledge a peripheral role for norms, in that speakers can take advantage of certain norms by flouting them. The cooperative principle, and the maxims of relevance, perspicuity, quantity and quality are norms that can be flouted to create conversational implicatures, where the speaker intends to mean (implicate) something in addition to the meaning of their utterance. He also

[51] Although I take this example from Davidson (1986), this view of language probably fits better with Davidson's earlier views on language; e.g. Davidson's (1967) Truth and Meaning.

[52] On Davidson's theory, Alexander has to appeal to what Davidson calls a "passing theory" of meaning for interpreting Diogenes' utterance. As I will explain in Chapter Four, this is principally to allow for occasions where what the speaker means is not what their statement typically means in the language. Since this isn't such an occasion, I'm ignoring this detail for the present.

acknowledges norms in the ways the utterances are correlated with the kind of response that the speaker intends the audience to produce(Grice 1957). But for Grice, norms are peripheral; at center-stage are the speaker's intentions and the meanings that are derived from these intentions (utterer's meaning, utterance occasion meaning, and utterance timeless meaning). The central role of norms, distinguishes me from Grice and other second-base theorists, placing me closer to Haugeland's "third base" theorists for whom such norms are a central source of intentionality.

3.2 *The intentionality of mental states: a problem for first and second base accounts.*

First base and second base approaches take an entity's having cognitive states with intentionality as the criterion both for whether the noises that the entity makes have intentionality, and for the particular intentional properties of that utterance (what objects or states of affairs the entity's utterances is directed at). This incurs a problem, in that this simply moves the mystery back one step. One mysterious thing (the intentionality of language) is explained in terms of yet another mysterious thing. We are still owed an account of the intentionality of people's cognitive states. In particular, we need a criterion for whether the entity really has them or not, and one for determining what particular intentional properties each cognitive state or item has. Until we have such an account, using the intentionality of people's cognitive states as the criterion by which we are to tell whether a noise or an ink-mark has intentionality is nothing more than a promissory note.

So far, however, we don't have an account that succeeds in fulfilling this promise. There allegedly is a fact, for most theorists, as to whether or not the entity in question has intentionality. This fact will determine whether the noises made by that entity have intentionality (meaning) and are anything other than just noises. For some theorists, only humans have intentionality. For others, humans and some other animals have it. For others, it's a graded notion; humans have lots of it, dogs have a bit less, snakes have very little, and ants have hardly any at all. We are fairly confident that human beings' neurological states have intentionality, however, and toaster-ovens internal states do not. But even of human beings this is based on little more than an intuition that *of course* human beings' cognitive states have intentionality. There is still much dispute over what makes it *true* that a person's cognitive state has intentionality. The problem is particularly prominent in discussions of non-humans: artificial intelligence, extraterrestrials and animal cognition. There is no established criterion by which to determine whether an entity, or an internal state of that entity, does indeed have intentionality.

There is even more dispute over what gives intentional states the particular intentional properties they allegedly have, as shown by the interminable debate on the problem of giving a naturalistic explanation for a representation's content in a way that allows for misrepresentation.[53] It is especially difficult to determine, for those entities that are unproblematically held to have intentionality, which objects or states of affairs their intentional states are about, such that when applied to something other than this (type of) object, they misrepresent that object. The problem is that this content has a normative component, in that each representation misrepresents when applied to things other than what it *should* represent. It amounts to the problem of giving a physical account of a normative property. Many naturalistic explanations of content have been proposed, but of all proposed so far, there are serious, well-founded criticisms. (I outlined a principal one –Fodor's disjunction problem—in Chapter Two.)

3.3 *Third base approaches: public symbols' intentionality is "original"*

First base, second base and third base theorists all mark some distinction between human beings and most other entities, in terms of humans' (superior[54]) ability to appreciate the difference between what we should do and what we can do.[55] But there is an important difference between first and second base theorists on the one hand, and third base theorists on the other, with respect to the question: "In terms of what do we specify what an agent should do?" To first and second base theorists,[56] a human being is a *rational* agent. So what an

[53] For recent overviews of the many attempts to solve this problem, see: Cummins (1989), Fodor (1990), Sterelny (1990), and von Eckardt (1993).

[54] I bracket "superior" here to mark the distinction between theorists who believe that humans are more rational or more able to follow norms, and those who believe that animals have no rationality and are simply machines that act by virtue of the disposition of their organs. (Descartes is a good example of someone who holds the latter; e.g. book V of A Discourse on the Method of Rightly Conducting the Reason. Aristotle is another.)

[55] The following contrast is made explicitly by Okrent (1996) §35ff.

[56] Haugeland says that for first base theorists, a human being is essentially a thinker; something that thinks rationally. Thus what the agent should do could be better phrased in terms of what the agent should will to do; the actual doing is irrelevant. For

agent should do is specified in terms of what it would be rational for the agent to do, given the agent's beliefs and desires, and by the nature of the physical environment the agent acts within. I should do whatever it's rational for me to do; i.e. whatever I believe will satisfy my desires. For example, if I desire that the light be on, and believe that flipping the switch will turn the light on (and have no other reasons for incompatible actions), then I should flip the switch. Here "rational" has some measure of normative force: I should do what it is rational for me to do.

For second and third base theorists, our ability to get along with one another is based in a large part of shared knowledge and a common foundation of rationality. By dealing with the same world, we come to have similar beliefs and similar knowledge to one another. This enables us to make sense of one another as rational agents, acting on their beliefs according to their desires.

Third base theorists concentrate, not on a common set of beliefs or knowledge about a common world, but on the shared practices of agents that interact socially within their world. Our ability to deal with our world successfully does not rely on a foundation of shared beliefs, or of rational deliberations, but on shared ways of acting, and judging—shared practices that we participate in together. Dreyfus argues this point forcefully (p. 142 ff.), saying that

> Heidegger's basic point is that the background familiarity that underlies all coping and all intentional states is not a plurality of subjective belief systems including beliefs about each others' beliefs, but rather an agreement in ways of acting and judging into which human beings, by the time they have Dasein in them, are "always already" socialized (Dreyfus 1991, p. 144).

This socialization ensures the basic agreement in judgements, or "agreement in form of life" (Wittgenstein 1958, §241), that is for Heidegger the basic condition for the possibility of any intentionality (Dasein).

Third-base approaches (which Haugeland (1990, p. 412-3) refers to as "neo-pragmatist"), see a human being as a member of a conforming community (Haugeland 1990, p. 417). What an agent should do is specified in terms of what it is appropriate for the agent to do *socially*, given the community's norms, rather than what it is appropriate for the agent to do *rationally*, given the agent's goals.

the present contrast with third base theorists, however, this distinction is not terribly important.

Here, the normativity that makes an agent's action goal-directed is external, not internal (see Okrent 1996, §37 ff.). The agent's ability to act as it *should* act rather than simply acting as it *can* act is the factor that distinguishes intentional action from mere behaviour.[57] What for second base and first base theorists are, respectively, objective principles of rational behaviour and of rational reasoning, are here normative, constitutive rules about what it is to be rational. Something *counts as* rational just in case it conforms to the community's standards of rational behaviour.

Similarly, for third base theorists, something counts as having intentionality, if what it is doing *counts as* acting purposefully and rationally, in a goal-directed fashion, following the norms of the society in which it acts and in which judgements about the rationality and intentionality of its actions are made. Its this condition of acting in a goal directed fashion according to the norms of your community, that Heidegger (1927/1962) refers to as being-in-the-world. As Okrent (1996)explains:

> "...the necessary conditions on the possibility of describing an agent as skillfully coping with her environment while following social norms, whatever those conditions might be, are at the same time the necessary conditions on that agent having any intentions whatsoever. That is, nothing can think unless it is being-in-the-world. (§57)

One of the aims of this chapter is to present my views on the nature of these necessary conditions on the possibility of describing an agent as having intentions. Here I agree with Heidegger in maintaining that these conditions are the same as the necessary conditions on the possibility of describing the agent as skillfully coping with her environment while following social norms. However,

[57] Okrent (1996) uses this point to argue that behaving in an en-minded way (acting purposefully according to the norms of the society in which you act) is the criterion for having a mind. And thus the internal constitution of the entity is largely irrelevant. Thus, he concludes, although it might be unlikely (§72), it is not impossible for a well-programmed computer to be Dasein:

> Precisely insofar as the two descriptions, 'acting purposefully as one should given a set of social practices' and 'acting in accordance with a set of rules for manipulating formal symbols,' are logically independent of one another, the behavior of some agent might satisfy both descriptions. The only way we could ever find out whether we could build such an entity (or, maybe, even be one) is to try to build one and see (or to try to come up with a set of rules which adequately captures our own behavior, which, under another description, is also skillfully coping with our environment, by acting in accordance with social practices) (§71).

my approach differs from Heidegger's approach in terms of the precise nature of these conditions. My approach also differs from these approaches in the kinds of entities that we take to be the primary bearers of intentionality. This will become clearer later on. For now we should get the general third base perspective on intentionality in general laid out.

For third base approaches, as described by Haugeland (1990), all intentionality is supported by and derived from the norm-governed way of life of which the bearers of intentionality are a part. These approaches solve –or rather avoid– the problem of accounting for the intentionality of cognitive states by inverting first and second base theorists' picture. They make the intentionality of public symbols –tools, actions and the symbols of public languages– primary, and derive other forms of intentionality –including the intentionality of people's cognitive states– from this. Haugeland situates himself, along with Heidegger, Sellars, Brandom, Dewey, Dreyfus, Dummett and McDowell together at third base (note 22). As will be apparent from the preceding chapters, I consider myself to be most "at home" at third base as well.[58]

Here the normative relationships are primary. They do not have to be explained away, in terms of non-normative physical properties, as most physicalist approaches require. Third base approaches do presume that there can be a naturalizable account of how norms arise in a community, but do not attempt to *reduce* these norms to non-normative relations and properties. Haugeland (1990, p. 404) gives the beginnings of such an account. Giving such an account –in terms of how such norms evolve and are preserved in a community's practices– is the goal of my final chapter.

For third base theorists, the norms of a community are *not* derived from the intentional states of individual agents. A community's norms exist as shared

[58] Haugeland (1990) says little about Wittgenstein. His only comment, comprising the entire section, is "Wittgenstein might have been a shortstop" (p. 403). Since the original inspiration for my position comes from Wittgenstein, I originally considered that my position fits between there, with a little from second base and a lot from third base, but distinct from either. This is because I was taking Heidegger as the paradigm example of third-base theorists. It is only in the last little while, as I came towards completion of this thesis that I realized how much myself, and Wittgenstein, have in common with third base theorists. Now I understand the third base position a little better, I believe that although Wittgenstein might have been a shortstop, his focus on forms of life as sets of distinctions, judgements and actions as set within practices, with shared criteria for following the rules of a practice correctly means that he would play well at third base too.

public property, as *the community's* norms, not as an aggregate of the intentions of the individual members of the community. An aggregative[59] account of norms is based in individualist, often reductive, accounts of cultural phenomena. Such accounts reduce cultural phenomena to the aggregation of the social actions of individual members of the society.[60] Thus the derivation goes in the reverse direction. The intentional states of individuals are derived from the community's normative practice of attributing intentional states one another, to explain and predict (and to induce conformity in) one another's behaviour. This is a theme that will recur again in the final chapter, when I talk about the properties of groups, and now the forces of natural selection can select in favour of groups who possess a certain property. I'll be using this principle to argue that norms can evolve through this process, by virtue of *the groups* having a certain practice (e.g. that one should give warning calls when one sees a predator). This is one of the principal disagreements I have with Brandom (1994). He argues that cultural phenomena, practices in particular, exist at the *I-thou* level, not at the *I-we* level[61]. They are grounded in the interactions between individuals, for Brandom. As I'll argue in Chapter Six, however, we can only understand the mechanisms by which norms arise (evolutionarily speaking) if we see practices as properties of groups, and only derivatively as properties of individual members. As properties of communities, any individual member of the community could fail to participate in the practice, as long as it is the community's norm that one should and does participate (and that one punishes those who transgress, so that individuals who transgress should be punished, even if they are not caught). The practice of ensuring conformity to the practices that the community's norms stipulate, (e.g. giving encouragement to people who behave as one ought to, labelling certain types of behaviour "irrational" or "immoral", giving examples of the correct way to do things, law-suits, jail, banishment, or even execution) ensure that most people conform to the practice, while also allowing the community's practices to evolve.

[59] The term "aggregative" is from Robert Wilson (1997b), following Sober (1991).

[60] Examples include most sociobiology (e.g. E. O. Wilson 1980), many accounts of the evolution of cultural phenomena (e.g. Sperber's (1996) methodological individualism (motivated by a fear of reifying "cultural beliefs", but ignoring cultural practices), and quite a bit of evolutionary psychology (e.g. many of the entries in Barkow, Cosmides and Tooby 1992).

[61] See Brandom (1994), pp. 38-39, and the entries under I-Thou/I-we sociality, in the index.

One of the central points of this chapter is to explain how the intentionality of people's cognitive states can be derived from shared public practices. My aim is also to adapt the third base theorists' account of what things count as the "public symbols" that have original intentionality.

Heidegger's approach is typical of third-base theorists. Here, as Haugeland (1990) explains, the primary bearers of semantic properties are "semantically articulated symbols" (p. 412). But these symbols are not the internal cognitive states of agents, but shared social symbols that are external to any cognitive agent. To Heidegger all intentionality —that of tools, of actions, of mental states, and of the tokens of public symbol-systems— arises by virtue of the relationships of these entities to and within a set of socially situated norms. The most basic intentionality to Heidegger, is the intentionality embedded in a "background"; a complex network of public objects, distinctions, norms, practices, relationships, and especially linguistic symbols that constitute what Wittgenstein (1958) calls a "form of life".[62]

The roles that these items and practices items play in people's shared forms of life define them as the kinds of entities they are, and confer on them whatever intentionality they possess (Haugeland 1990, p. 413).

> "The idea is that contentful tokens, like ritual objects, customary performances, and tools, occupy determinate niches within the social fabric—and these niches 'define' them as what they are. Only in virtue of such culturally instituted roles can tokens have contents at all" (Haugeland 1990, p. 404).

Tools are a good example of such "contentful tokens". For example, a hammer counts as a hammer by virtue of the norms governing the use of objects like that to hammer in nails. Without these norms, it is just a piece of wood attached to a heavyish piece of metal with a flattened end. Dreyfus (1991, p. 63) argues that a piece of equipment "is *defined* in terms of what one uses it for" (where Dreyfus uses "one" to refer to a normal member of a community, doing what *one* normally does). But a tool is not defined in terms of "what one uses it for" in a narrow sense, but in terms of how it fits in with an overall equipmental nexus. What an object is "is its place in a context of use" (Dreyfus 1991, p. 64). The hammer is *for* driving in nails by virtue of these norms governing the practice of

[62] The use of the term "form of life" here is mine, not Haugeland's. Dreyfus (1991, p. 144), however, explicitly states that he interprets Heidegger's talk of background practices in terms of a Wittgensteinian "form of life".

driving in nails, and of building things out of wood, etc., and the norms governing the construction of hammers.

Words are a subset of tools on this account. So, similarly, a vocal utterance of the word "promise" is a token of a type of sound that fits into a certain norm-governed practice. This sound counts as an utterance of the word "promise," rather than just being a meaningless noise, because of the norms governing the use of that word when participating in the practice of promising. The word is *for* making promises, just as a hammer is *for* hammering in nails. (I'll have much more to say about the way norms define things as the things they are later on.)

These shared norms and the shared practices (judgements, forms of life[63]) within which they are embedded are, for third base theorists, a precondition for any individual intentionality. So the ability to act skillfully, according to social norms —the ability to act the way one should act, because that's how one should act— is a necessary characteristic of beings with intentionality. Okrent (1996) argues that for Heidegger, acting according to norms of how one should act is the mark of an intentional agent:

> One is Dasein (has intentions) only if one acts, one acts only if one is in-the-world, and one is in-the-world only if one acts skillfully according to the practices of the society in which one lives. (§56)

An agent, to Heidegger, has intentionality (is Dasein) only by virtue of the agent's ability to skillfully cope with its world by acting within a context of social norms and practices.

Thus, on the third base view, it is only when people are socialized into a form of life, that they can have any intentions at all. Babies get socialized, says Dreyfus (1991, p. 143), but they do not have intentions –"they do not Dasein (verb)," he says– until they are already socialized. For Heidegger, we become Dasein when we begin to act within a world that is shared with others.

> "By "others" we do not mean everyone else but me—those over against whom the "I" stands out. They are rather those from whom, for the most part, one does not distinguish oneself —those among whom one is too.... By reason of this *with-like* being-in-the-world, the world is always one that I share with others. The world

[63] C.f. Wittgenstein (1958), §241-2.

> of Dasein is a *with-world.* Being-in is *being-with* others. (Heidegger 1927/1962, p. 154-5)

So the primary way of being here is not as a self, an "I" (especially not a Cartesian self, where the problem of other minds appears), but as part of a "we"; a community of like-minded entities who agree in form of life.

Thus the ways in which we are socialized into being-with-objects (using tools, for instance) are *shared* ways of being-with-objects. In other words, the norms that specify tools as the kinds of tools they are, and that specify what they are for, are shared norms. We are socialized into seeing these as appropriate ways of using such objects.

Because of these shared norms that permeate and make possible our forms of life and our intentionality, there is an important distinction here, between how a tool *can* be used and how one *should* use it. For example, Haugeland (1990) points out (p. 409) that, in spite of the fact that a screwdriver can be used to prod laboratory rats and carve one's name in picnic tables, one should use a screwdriver for turning screws. It's that the screwdriver is *for* turning screws, according to the norms that screwdrivers function within, that makes it a screwdriver. (Thus a broken screwdriver is still a screwdriver because of its role in this normative structure, in spite of the fact that it can't be used as one should use a screwdrivers.) By virtue of its shape and availability in workshops it also has "derivative" uses, such as for opening paint cans. But using a screwdriver to carve initials in a picnic table would be a mis-use of the screwdriver. This is not what screwdrivers are for, this is not how one should use a screwdriver. Similarly, using a chisel to turn screws would be a mis-use of the chisel. Such uses are often justifiably subject to censure; a workshop supervisor once berated me for using my chisel in just this way. (Compare promising to do something, in order to secure someone's cooperation in a task, while having no intention of doing as you promise. This would be mis-use of the word "promise" and justifiably subject to censure.)

These norms of how tools should be used are not isolated rules about particular tools, for Heidegger, but are part of a large set of normative "equipmental" relations between objects. Tools (or "equipment"), according to Dreyfus (p. 62), are set within an interlocking interdependent network of relations between equipment that constitute (in part) the shared form of life within which we live and act. In Heidegger's words:

> Equipment—in accordance with its equipmentality—always is *in terms of* its belonging to other equipment: inkstand, pen, ink, paper, blotting pad, table, lamp, furniture, windows, door, room. (Heidegger 1927/1962, p. 97)

Screwdrivers are for turning screws; this is what it is to be a screwdriver. Screws are for fastening things to wood or metal. This is contrasted with situations where nails would be more suitable (should be used instead) for holding things to wood. Hammers are for forcing nails into wood. Nailing or screwing things to bits of wood is for building furniture and houses. Furniture is for sitting on, putting things on, putting things in. Furniture goes in houses. Houses are for living in. And so on....

Haugeland argues that this kind of complex and highly interdependent structure of normative relations among public paraphernalia is "the third-base archetype of all intentionality" (1990 p. 409). The public paraphernalia that we skillfully cope with are the primary possessors of intentionality. They get this intentionality (e.g. a hammer is revealed to my behaviour as being directed at nails that should be hammered into wood that should be fixed to other wood...) by virtue of this normative nexus. I will discuss the kind of intentionality we have when we skillfully cope with objects later on. For now, however, it's important to note that it is derived from this normative context of equipmentality. My "comportment" (i.e. my behaviour) reveals objects, in their roles in this nexus of equipmentality. My desk, for instance (see Dreyfus 1991, p. 68), is revealed to my comportment as a surface for writing on, or a place to keep things, depending on the activity I am pursuing (the practice I am participating in). The desk has this use (this *being in-order-to-write-on*), however, not simply because I skillfully cope with it by writing on it, but because the equipmental nexus that the desk and I exist and act within includes the norm that desks are for writing on.

3.4 *A background of tacit norms*

When we skillfully cope with objects, the nexus of equipmentality relationships that define the objects we manipulate, and that specify what they should be used for, are often not consciously entertained nor explicitly formulated. As Dreyfus (1991, p. 4) explains, Heidegger denies that most of our everyday understanding is explicit. He also questions the possibility and the necessity of making it explicit. Heidegger introduces the idea that

> the shared everyday skills, discriminations and practices into which we are socialized provide the conditions necessary for people to pick out objects, to understand themselves as subjects, and generally, to make sense of the world and of their lives (Dreyfus 1991, p. 4).

However, Heidegger argues that these practices primarily and usually function only in the background. They operate tacitly, constraining the way we see our worlds and the way the world "demands" certain actions of people skilled at operating within a particular context. The way we use a knife and fork for eating with, while we concentrate on the dinner-table conversation, is a good example. While dining, using your fork to pick up the food from your plate and bring it to your mouth is just what the situation demands of someone skilled and familiar with the practice of eating with a fork.[64] Thus in many situations, actions the norms deem appropriate are simply "what one does", what the situation demands. We ordinarily wouldn't consider the possibility of acting differently.

Human beings' norm-following behaviour, however, has some unique features. We can engage in critical reflection on our norms, and on the context of equipmentality they bring about. Heidegger's position, however, is that such reflection is only possible because of a large amount of inexplicit and unarticulable skills, discriminations and practices into which we have been socialized and about which we do not normally reflect. Moreover, alternative actions only become apparent in cases of "breakdown", when a tool is broken or missing. For example, ordinarily when the door needs stopping open, the wedge-shaped piece of wood on the floor behind the door becomes apparent as a doorstop. Using the doorstop to stop the door is simply what one does. It is what one does unthinkingly, habitually –in the same way we unthinkingly use the knife and fork while engaged in the practice of eating at the dinner table–without the need for any reflection on what you are doing or why this is the right thing to do. The norms that govern how the tool should be used, in ordinary cases where the tools are ready-to-hand and the agent is using them while engaged in skillfully coping with its world, are not entertained by the agent. They appear to the agent as appropriate ways to use the object. They are

[64] Also consider the ways in which we recognize and produce facial expressions and "body language" indicative or one's reaction to an event: impatience, surprise, interest, concern, delight, alarm, disappointment, and so on. Our following norms of social interaction with others depends to a large degree on shared ways of non-linguistically expressing our reactions in this way.

a part of the largely tacit "know-how" of a socialized skilled agent, not as explicit "knowing-that" style rules about how the object should be used. I know how to use the doorstop to keep the door open, I do not "know that" doorstops are for keeping doors open; at least not while I am engaged in stopping the door open. The situation, in this view, "demands" a certain move from me; a move that I can make quite unthinkingly. Similarly, a norm like "Forks are for skewering bits of food onto, in order to convey them to one's mouth", is rarely explicitly entertained by people while they are engaged in the practice of eating with a fork. (For defense of this position, see Dreyfus (1991), p. 85 ff.) I *know how* to use a fork, and when one is available I use it as the situations demands.

In cases of "breakdowns", where the appropriate tool is broken or missing, then we think about our goals and become conscious of alternative ways of achieving the for-the-sake-of-which of the unavailable tool (see Dreyfus 1991, p. 71 ff.). Trivially, if I sit down to eat and do not find a fork, the absence of a fork is noticeable in a way that the presence of a fork would not be. The situation demands the use of forks, and none is available. The absence of the fork that I would normally unthinkingly pick up and use, requires me to think about what to do, and make a conscious decision abut how to meet my goal of eating the meal. I might look for chopsticks instead, if the situation is such that the use of chopsticks might be proper or expectable. Alternatively, I will ask for a fork, or go to the kitchen and get one. In my other example, when there is no doorstop and I need to keep the door open, for instance, then I do think consciously about the properties of the doorstop that make it *for* stopping the door (being wedge-shaped and not too smooth and slippery), or the properties of other available objects that make them suitable for using to keep the door open (e.g. solidity and being immobile enough that the door won't be able to move it). In such cases I would consider either making a wedge out of something else (a folded piece of paper, perhaps), or using something heavy (a large book, or a heavy toolbox). Recently someone was doing some work around the humanities building where I write, and needed to stop the door open. I observed him open the door and immediately take two screwdrivers from his tool-belt, cross them, and use the upper one's tip as a wedge under the door, its handle end being propped up by the shaft of the other screwdriver. This appeared to be a "normal" procedure for this person. He needed a doorstop, and in this context his screwdrivers became apparent to him and good for stopping the door. For him, this could have been so familiar and often-employed solution to the problem of the lack of a doorstop, that the situation wasn't even perceived as one in which a doorstop was lacking. This could have been a background, almost tacit, use of the "doorstop" he carries in his tool-belt, as demanded by the situation. This could, however, have been an ingenious invention, and figured out by deliberating on the properties that make doorstops good for stopping doors, and realizing that a pair of screwdrivers could be good for that task as well.

Dreyfus and Dreyfus (1982) present an account of skill-learning that works in the reverse direction. We learn explicit rules that are consciously followed, noticing the properties and relations that underlie the skill in question. After a large amount of practice, this skill becomes part of our background of skills, as we can perform elements of the task rather unconsciously, focussing more on the goal of the task than on the means by which we achieve it, and the tools that are part of those means. When driving, for instance, we initially concentrate on pressing the brake-pedal with the right "touch", turning the wheel, shifting the gears in time with depressing the clutch. After some experience we concentrate on merging with the traffic, avoiding the potholes, and passing slow cars.

Michael Polanyi (1958, p. 55 ff.) also presents an account of skill learning in which he argues that our ability to focus consciously –or in *focal awareness* as he calls it– depends upon a similar set of background familiarities and skills. In learning a skill –for example, a person learning to hammer in a nail, or a blind person learning to use a cane–, in which the agent's focal awareness is initially on the particular movements of the cane or hammer in the hand and how to interpret them. However, as the agent becomes more experienced, the particular movements of the cane or hammer in the hand eventually move into what Polanyi calls *subsidiary awareness*. The agent's focal awareness moves to the overall practice and its objectives, such as walking while feeling objects and surfaces with the tip of the cane, or making the nail flush by hitting the nail with the head of the hammer. The movements of the cane or hammer in the hand are part of the person's experience, but a part that the agent does not concentrate their attention upon. The exercise of using the cane to focally attend to objects in one's path, then, depends upon the exercise (in subsidiary awareness only) of a number of other skills.

John Searle (1995) has a similar analysis of "the background". To Searle the thesis of the background is that: "Intentional states function only given a set of Background capacities that do not themselves consist in intentional phenomena." (p. 129). These background capacities are a set of abilities, dispositions, tendencies and *causal structures generally*" (p. 129) These background capacities allow us to develop skills and dispositions that are responsive to institutions and their norms, without consciously or unconsciously applying those norms. In many cases, because of the way we have been socialized and trained to act according to the norms of the social institutions we act within, we have skills that enable us to respond appropriately to situations without actually representing those rules to ourselves. The norms are neither consciously nor unconsciously entertained as we act according to them, "we just know how to deal with the situation" (p. 143).

The point, here, is that in all our conscious activities depend upon a large amount of tacit skills and know-how. Even conscious deliberation on the properties of a tool that make it good for what it is for, is only possible against a background of many other unarticulated and unarticulable background practices, skills and discriminations that do not operate at the level of conscious deliberation.

3.5 Most animals operate principally at the level of the background

Heidegger distinguishes creatures that have Dasein (socialized human beings) from the animal kingdom based on this ability to skillfully cope with one's world by using tools as one should use them, as opposed to only being able to use them as they can be used (Okrent 1996). We can express this in terms introduced by ecological psychologist J. J. Gibson (1979). To Gibson, an animal's "niche" is partly constituted by "opportunities for actions" or "affordances". Objects are presented to animals in terms of basic uses they afford that (kind of) animal. For instance a tree affords climbing to an ape, affords nest-building to a bird and affords claw-sharpening to a lion. Heidegger sees human beings as building on this basic level of being-with-objects, by skillfully coping with objects within a nexus of shared practices that defines those objects not only in terms of how they *can* be used (their affordances) but in terms of their normative equipmental relations: how they *should* be used, or how one uses them.[65] Animals, to Haugeland (1982, 1990), do not use tools within such a normative framework,

[65] Ulric Neisser (1989) extends Gibson's account, however, to include cultural affordances afforded to humans by objects; for example the way mailbox affords letter-mailing. These depend upon a culture's norm-governed practices.

such that the issue of "propriety" comes up for them. Haugeland argues that even if an ape is skilled at getting bananas by using a stick, "an ape could not misuse a stick, no matter what it did" (1982, p. 18). On this view the ape could not misuse a hammer or a screwdriver, either, since Haugeland assumes that animals are not "socialized" into a normative framework where there is a distinction, for the animal, between how the hammer, screwdriver or stick can be used and how one should use them. Dreyfus (1991, p. 62) makes similar points about the context of equipmentality being absent in an animals use of an object. While a stick may be used to knock down bananas, the stick is not *for* knocking down bananas. In order to count as *being- in-order-to* knock down bananas, the stick would have to be a token of a type of tool, that is part of a nexus of equipmental relations.[66]

This difference between humans and animals, however, might better be characterized as a difference in the *degree and style* of being-with-objects, and of being-with-others, that each is capable of. Most non-mammals, at least, it is probably safe to assume, operate completely at the level of the background, without the added benefit of humans' often conscious engagement with their worlds. They operate by virtue of non-conscious "capacities, dispositions and causal structures generally" (as Searle (1995, p. 129) describes the background), at the level of engagement with an "available" world as it is presented to the creature (see Chapter Two). Most creatures' lack of abilities to solve "representation hungry" problems –their lack of ability to act with respect to absent, counterfactual or abstract features of their worlds– reinforces this idea.

However, there are norms within such worlds that –very tacitly– specify what the creature should do, and what it should do with the objects presented to it. The leaf, to an ant, affords food, and should be brought back to the nest. Small sticks afford nest-building to birds; a bird that, while engaged in collecting

[66] Recall here Peirce's distinction between icons and indices on the one hand, and symbols on the other. Icons and indices can function singly, and be interpreted as signs of certain objects. But to be a symbol, is to be part of a system of interpretation, where there are relations between the symbols, based on the relations between their interpretants. Heidegger requires that equipment is part of a system of equipmental relations between general types of tools (of which each particular tool is a token) as a condition for intentionality. Peirce requires that symbolic systems instantiate relations between general types of symbols (of which each use is a token) as a condition for a sign having content. This points towards interesting connections between the two, that might be worth exploring. Exploring these connections would be beyond the scope of this work, however.

sticks and building a nest, identified perfect nest-building sticks, but ignored them, could *perhaps* be said to be "mis-using" the sticks.

But although we humans might describe this sort of situation as acting inappropriately, the more important issue is whether the creature's conspecifics would judge this to be so. There is an important distinction to be made here, between whether such norm-describable behaviour is due to genetic programming, or due to socialization. Did the rules cause the disposition to act in this way during the creature's lifetime by behavioural conditioning in the creature's *social* environment, or did they develop based on pre-programmed genetic building blocks, perhaps conditioned in response to conditioning from the *physical* environment? Situations where the creatures' genetic "programming" make it the case that they are caused to do these things do not count as normative. These are causal mechanisms. Such creatures cannot act otherwise. Creatures that do act otherwise could be compared to broken machines; it's causally impossible (by definition) for a *properly* functioning machine to fail to do as it should. The norm here (the ant should bring the food back to the nest, the frog should only snap at flies) is attributed only by people, who have an understanding of the mechanism and what it is good for (selected for). It is not a norm for the creature and its conspecifics. (Compare Millikan's (e.g. 1989) notion of Proper Functions that, through people's knowledge of the mechanisms of natural selection and what the mechanism was selected for, determines what a biological mechanism in an animal "should" do.)

Many ethologists refer to the case of a particular wasp that brings food to the entrance to its nest, goes inside to check the nest, and then comes out to bring the food in. If the food is moved a couple of inches away, the wasp will bring it back to the entrance, and then leave it there again while it checks the nest. This behaviour will be repeated endlessly, as long as the food is moved while the wasp is checking the nest. The norm that the wasp should check the nest before bringing the food in is imputed to the wasp by human observers. This is genetically programmed, rather than normatively enforced. The wasp appears incapable of defying the rule. It's a causal regularity rather than a normative rule.

Only socially conditioned patterns of behaviour, where the behaviour is enforced by a creature's community, should count as norm-governed, as opposed to being causal regularities. Only when there is socialization —enforcing of the norm, by censuring behaviour that contravenes the norm, and rewarding behaviour that conforms to it— should situations count as situations where norms of behaviour and of the proper uses of objects govern creatures' behaviour. Here the norm is responsible for bringing the behaviour about. And here behaviour that is wrong is wrong not simply because it deviates from the rule (as a broken machine or creature could) but because the behaviour is subject to censure. The practice of censure is what specifies the norm: what kinds of behaviour *count as* (by being *treated* as) according with the norm and what kinds of behaviour *count as* (by being *treated* as) contravening it.

Wittgenstein (1958) points out that a community's judgements about behaviour establish the precise content of a norm. Here he addresses the problem that on any interpretation of the content of a rule, a different content could be interpreted. And counterfactual cases that on one interpretation contravene the rule, could be interpreted by the other to accord with the rule. Wittgenstein's solution is to point out that "...there is a way of grasping a rule which is *not* an *interpretation,* but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases" (§201). The rule is grasped not in having an explicit interpretation of the content of the rule, but in our treating certain behaviour as appropriate. Following Wittgenstein, Brandom (1994, p. 63) makes similar points about how to determine the content of a norm, by looking at the actual *practice* of assessing actions. These assessments are not to be understood as based on propositionally explicit beliefs or commitments about the content of a rule, because then the regress problem recurs, in that the application of the explicitly stated rule is also subject to interpretation. The solution, says Brandom, is to look at the actual practice of applying sanctions (behaviour that positively or negatively reinforces the behaviour responded to).

Socialization of an animal –often through the mechanisms of behavioural conditioning by its conspecifics– "teaches" the animal how it should behave, how an object should be used, and so on. Such socialization for tool use does happen in the animal kingdom. There are plenty of examples of animals following socialized norms for the appropriate use of objects, and of animals being socialized into participating in social practices. Norms of appropriate behaviour, for instance, constrain the behaviour of members of herds of cattle or horses; for example Monty Roberts (1997), the "horse whisperer" employs horses' methods of censoring aberrant behaviour (the matriarch of a herd "excludes" the offender for a while) to train horses.

There are even socialized uses of objects in the animal kingdom. Tool using behaviour is very prevalent in the animal kingdom, from insects (ants "farming" fungus) and birds (crows throwing stones at eggs to break them), to rodents and elephants (elephants use all kinds of sticks and swatches as tools, mostly for personal grooming). It's in monkeys and in great apes that the most sophisticated tool use, and the social transmission of tool use techniques is especially prevalent.[67] Chimpanzees use of "fishing sticks" to get termites out of their underground homes (e.g. McGrew and Rogers 1983) is a good example of socialized tool use. Chimpanzees "fish" for termites by thrusting the stick deep inside the termite colony, waiting for the termites to attack the stick, and then quickly pulling the stick out of the colony while it is covered in (presumably nutritious) termites for the chimpanzee to eat. This is only practiced in the wild by certain troops of chimpanzees, however. Some groups of chimpanzees live their lives in places where termites are also in the area, but because they lack this practice of fishing, they do not take advantage of this –to them unexplored– food source. Some groups of chimpanzees also use hammer stones to open nuts; the presence of this practice is also not universal; many chimpanzee troops live near sources of nuts but do not know how to take advantage of this food source. Groups of chimpanzees where one member learns or knows how to use a hammer-stone and anvil to open nuts also rapidly pick up the practice. Hannah and McGrew (1987) reported on the spread of the technique of using a hammer stone through a population of chimpanzees released from captivity into the wild, where initially only one of the members of the group knew how to do this. Boesch and Boesch (1990) report that of the chimpanzee troops that do know how to open nuts, techniques for selecting and using hammer stones and techniques for making wooden tools vary between groups, but within groups the techniques and the results of tool manufacture are very similar. Furthermore, only some tribes (only one of three that Boesch and Boesch (1990) studied) extend this technique to using tools to extract the brains of monkey prey.

[67] I owe thanks to Michael Snyder for bringing this literature on animal tool use to my attention.

3.6 *Human intentional practices arise out of pre-conscious "background" capacities.*

Human practices and forms of life, with our capacity for conscious deliberation on (some of) our norms can be seen to arise out of such pre-conscious non-intentional norms. Humans have developed several remarkable additions to the ways of being-with-objects exhibited by animals. Rather than just doing what we have been socialized and conditioned into doing, we human beings can think about what we do and can make a conscious effort to evaluate what we do and to figure out better ways of doing things. We can think about the available objects, and what makes them good for what one uses them for. We can modify objects to make them better tools. We can even "decontextualize" our tools and reveal their context-free features, e.g. in the construction of scientific theories. (Dreyfus p. 84)

Humans' abilities to deal with "breakdown" cases by explicitly focussing on the properties of tools that make them "good for" certain tasks, arise out of the kind of pre-conscious background capacities that are the whole form of life for many animals. This also can be seen in the capacities of our closes relatives to transcend the humble ways of being of other animals.

In addition to being largely driven by the same background capacities and causal relationships that permeate all animal life, including human life, chimpanzees can also operate at the more conscious level. In addition to using a tool as it should be used, chimpanzees have demonstrated the ability to consider what about a tool makes it good for that use. For example, when using tools some chimpanzees have demonstrated the ability, when a tool is unavailable, to (in Dreyfus' words) "confront its equipment in context as somehow defective, and can try to fix or improve it and get going again" (1991, p. 84; Dreyfus is talking about humans, not chimpanzees here, though). For example, Sue Savage-Rumbaugh reports[68] that in an experiment to investigate the evolutionary origins of human tool use, the chimpanzee Kanzi was taught to make and use an edged stone tool to cut a rope to obtain food. Kanzi disappointed the researcher, however, by coming up with a novel innovation. Rather than striking two stones together (as presumably humans' ancestors did), Kanzi threw the stone at the hard tile floor of his enclosure. This technique struck off much larger shards (Kanzi is very strong and can throw much harder than an average human). The experimenter was not happy with this

[68] As cited by Andrew Fenton (2000).

development, however, and the next day the cage was covered in carpet in an attempt to prevent Kanzi from continuing with this technique. Savage-Rumbaugh reports that when Kanzi discovered that the tile floor was covered with carpet, Kanzi tried a few times unsuccessfully to break shards off the stone by throwing it, contemplated a while, and then searched for a join in the carpet. Kanzi pulled the carpet back from the join to reveal the tile floor, and threw the stone there.

Some chimpanzees also use tools in what might qualify as a context of equipmentality, for Heidegger. Brewer and McGrew (1990) report that a wild chimpanzee which they named Katie created a set of tools in order to get at honey in a bee-hive. The bees in question were small and stingless and had evolved so that they coated their hive in very hard batumen for protection. Katie made a set of tools, each for a specific purpose, oriented towards extracting honey. She first made a dipping probe, a long stick that she inserted into the hive to reach the honey. When she discovered that she couldn't reach the honey she made a chisel by breaking off the end of a dead tree. This was a rather strong tool with a sharp end, which she jabbed into the hive with overarm swings until she made the batumen start to crumble. Next Katie made a sharper and smaller chisel with which she carefully enlarged the hole in the batumen. She then made what Brewer and McGraw refer to as a bodkin, a green stick, 1 cm thick and 30 cm long, which she used to thrust through the internal seal of the hive. Finally another dipping probe was made that could be held in a "pencil grip" to extract honey from the hive's well. This set of tools was made, each with a different but interrelated purpose, in the pursuit of a goal.

In summary, then, for third base theorists, with whom I have a significant amount of agreement, there are two basic pre-conditions for the possibility of intentionality. The first is a set of norms governing how one ought to behave, how things ought to be used. The second is that members of a community *share* these norms. So far I have argued that the combination of these two conditions results in the further point that these norms are for the most part unarticulated; they undergird the shared "background" practices and distinctions of the members of a form of life. Because of this shared background, a community's members act and judge similarly, sharing a common sense of appropriate ways to behave and appropriate ways to use tools. But even if these shared practices are a precondition for intentionality, I still need to answer the question: what of intentionality itself? What kinds of things have intentionality (original or derived)? And how does this norm-governed form of life, with its shared corpus of background practices and distinctions, give rise to intentionality? We can begin to find answers to these questions by looking at a special sub-set of the tools people use; those linguistic tools that people use *together* with other people.

But while some higher apes may share some of the "background" capacities and some of the conscious capacities that enable human beings to do what we do, an essential difference still exists between humans and all other animals. This difference is that while norms can be transmitted and enforced socially by the practice of censure and encouragement, human beings are unique in that we have the abilities to attribute intentionality to one another. Some very primitive forms of non-verbal "mindreading" have been observed in chimpanzees (see Whiten (1993, 1996b) for a diagnosis of the limited extent to which our closest non-human relatives can mindread). While they can perhaps read what someone else sees (Whiten 1993, p. 376), they have trouble extending this to what they are aware of (p. 378) and can make judgements about what others might or might not notice, but not necessarily about what others might and might not know. There have been no documented cases of chimpanzees having any sensitivity to others' false beliefs. The one false belief experiment Whiten (1993) reports the chimpanzee failed to demonstrate sensitivity to the difference between what the chimp knows and what someone else believes (p. 378). Thus the ability to explain and predict others' behaviour on the basis of their intentional states remains firmly in the human domain.

This difference, as I'll argue soon, makes all the difference. It's evolving this practice that enabled languages to develop, such that we can make our norms explicit, and can make contentful ascriptions of intentional states within the practice of giving and asking for reasons. It's this set of abilities, then, that for third-base theorists is where intentionality is at home.

In the next chapter, I'm going to outline and defend a "third base" approach to language (by modifying a second-base Gricean account). This account will show the centrality for our linguistic abilities of the ability to attribute intentional states to others within shared social practices. The succeeding chapter will develop this point, showing how the felicity conditions on speech acts and the criteria by which acts are judged felicitous, are part of normative public practices. The intentionality of mental phenomena are shown also to emerge within linguistic practice, in that the same felicity conditions apply to the speech act of attributing intentional states to others and to oneself. The point of all this is to show the thoroughly normative nature of intentionality: that it only exists within and because of normative social and linguistic practices. The final chapter will present an account of how such practices, and the ability to adopt the intentional stance could have evolved, thus providing a naturalized but non-reductive account of intentionality.

Chapter Four

An Embodied Action Approach to Language-use.

An ounce of practice is worth a pound of precept

—an English Proverb

Once the child has learned the meaning of "why" and "because" he has become a fully paid-up member of the human race.

—Elaine Morgan (1995) [69]

4.1 *"Problematic" examples of language-use.*

Consider the following situations. One morning not too long ago, my partner had her coat on, her briefcase waiting by the door, and uttered to me in a frustrated voice, "I'm late! I have to go now, but I can't find my key-cars." I knew immediately that she couldn't find her car-keys, and that she had said this because she wanted me to help look for them. Another situation: my grandmother once countered my grandfather's rather convoluted directions, saying "But there's a simpler way to get there; if you go his way there's so many more turners to corn." We all knew she'd meant to say that there's too many corners to turn. Another example: My roommate and I made pizza for dinner recently. While she was sliding the pizza onto the cutting board to slice it, she said, "Can you grab the wheel-thingummy from the drawer?" I knew that she was talking about the sharp-bladed wheel for slicing pizza, which is kept in the drawer I was standing in front of. I also knew that she was not inquiring about my ability to grab the pizza-slicing wheel, but that she wanted me to grab it, and to hand it to her.

Sometimes speakers make slips of the tongue, malapropisms, use familiar expressions[70] in unfamiliar or ungrammatical ways, or use expressions we've

[69] Quoted in Dennett (1996), p. 153

[70] I use "expressions" here to cover both sides of a distinction that I do not feel a need to rule on: the distinction between whether it is a word or a sentence that is the fundamental meaningful unit of speech. Since for me the fundamental unit of "meaningful" speech is

never heard before. But often in such situations we can nonetheless understand what the speaker intended to say and can respond appropriately. Donald Davidson in his paper "A Nice Derangement of Epitaphs"[71] points out that the fact that we can grasp what someone intended to say in such cases has important ramifications for "standard" theories about what it is to know a language, including theories he himself is responsible for (p. 437). Because such theories view knowing a language as having a theory by which the meanings of utterances can be interpreted, they have difficulty explaining our ability to successfully interpret utterances where unfamiliar expressions are used, or where familiar expressions are used in unfamiliar or idiosyncratic ways. Davidson suggests modifications to such theories to account for this ability.[72]

I will argue that the modifications Davidson suggests are not in order. To understand most of language-users' abilities –including the ability to successfully deal with slips like those above– we require not a modification to the "standard" theories that Davidson talks about, but an entirely different view of what a language is, and what it is to know a language. These "standard" theories are based on a very over-simplified model of how people use language: a model of speakers uttering sentences that they hold to be true, and audiences interpreting the meanings of such sentences by employing a theory about the meanings of the expressions used and how they are combined.

When theorizing about a language *itself* –especially in the context of viewing the language as an abstract formal system— there may be some purpose to theorizing in abstract terms about words' meanings, the relationships between them, and how they each contribute to the meanings of combinations of words. These abstractions of the phenomena of interest are, on the face of it, to simplify the phenomena and extract the important aspects of the phenomena

the speech act, I use "expressions" to refer generally to either a single word or a group of words, or a sentence.

[71] Donald Davidson, (1986). All references to Davidson will be to this paper unless otherwise stated.

[72] It is interesting to note at the outset that J.L. Austin (1962) noticed "slips" like those I mentioned above; he called them "flaws" (p. 17). He also remarked on the trouble a theory of language that explains linguistic phenomena in terms of meanings can have in accounting for such utterances:

> Somebody 'says something he really did not mean'—uses the wrong word— says 'the cat is on the mat' when he really meant to say 'bat'. Other similar trivialities arise—or rather not entirely trivialities; because it is possible to discuss such utterances in terms of meaning as equivalent to sense and reference and get so confused about them, though they really are easy to understand. (pp. 137-8)

from the irrelevant aspects. But these supposedly important aspects —sentences and their meanings and their truth-conditions— are central only for certain theoretical and analytical purposes. For other theoretical and analytic purposes, they may not be central, and in such cases, viewing the phenomenon as though they are central can give rise to rather unwieldy, *ad hoc* explanations of people's linguistic abilities. My point in this chapter is that while this perspective may be suitable for analyzing the language itself, as a formal system, it is not well-suited to explaining most of people's complex, context-embedded, linguistic actions and interactions. And it is especially ill-suited to explaining the abilities that enable these interactions to take place. Davidson's explanation of our ability to successfully interpret the kind of "slips" I just mentioned, as I'll show presently, is a good example of such an unwieldy, *ad hoc* explanation.

I will argue here that it's more illuminating to view people's linguistic abilities as a subset of the abilities that enable us to participate in socially structured interactions, rather than viewing them as abilities to produce and interpret meanings. The "standard" theories' abstractions of linguistic phenomena, as the production and interpretation of meanings, abstract the sentence uttered away from many aspects crucially important for explaining the act of uttering the sentence in the performance of a speech act, and for explaining the audience's ability to respond appropriately. These aspects of the performance of a speech act and the audience's ability to respond appropriately include the physical context of the interaction, the social context and the practices being engaged in (e.g. in a grocery store, engaging in the practice of buying and selling goods). The social relationship between speaker and audience and the immediate and long-term history of their interactions is also relevant. They also include the interactive purpose the speaker intends to achieve through this act, the ends the audience comes to believe the speaker aims at achieving and the means the audience comes to believe the speaker aims at employing. All of these aspects of the interaction, and much more, can be crucial to analyzing the speaker's act of making the utterance and explaining the abilities that underlie its performance and interpretation. These aspects often need to be emphasized, not abstracted away.

Uttering a sentence whose meaning is a proposition that you hold to be true (or a proposition that you *mean*) is only a subset of linguistic phenomena, and one of minor importance. I will argue that it is certainly not the paradigm example of language-use, upon which we can base analyses of all linguistic phenomena. I see uttering a sentence that the speaker holds to be true as a specialized case; a refined subset of the broader phenomenon of people performing speech acts. Furthermore, I see people performing speech acts as a subset of people interacting with one another while participating in shared social practices. Thus knowing a language is not the possession of a theory of meaning, but having a set of abilities that enable the agent to participate in certain sorts of practices. I want to concentrate our attention on this more general phenomenon of interaction within a practice, with a focus on those interactions that involve the performance of speech acts, and the abilities that enable such interactions to take place. Here I will explain and employ analytic tools specifically designed for analyzing such practices, interactions and abilities. I believe these tools are more illuminating and analytically useful –especially when talking about people's linguistic abilities– than using or extending (over-generalizing) tools designed for analyzing the more specialized and abstracted (over-simplified) phenomenon of people uttering sentences that they hold to be true, and people interpreting such sentences according to a theory of the meanings of the words, and of the systematic ways they contribute to the meanings of sentences.

Because uttering a sentence that the speaker believes to be true is *not* the primary example of language-use, mischaracterizations of people's everyday linguistic interactions and linguistic abilities can result from analyzing such interactions and abilities in terms of the sentences uttered and their intended meanings and truth-conditions. Davidson's (1986) account of the skills and knowledge that support people's ability to converse with one another is a good example of this kind of mischaracterization. It is common to analyze language-use in terms of what Davidson calls "first meaning" —the meaning a person intends their utterance to be interpreted as having (p. 434-6)– and the truth conditions of those sentences. Davidson supplies us with an example of such an analysis, when he analyses the example of Diogenes saying to Alexander the Great the Greek equivalent of, "I would have you stand from between me and the Sun." Davidson says that Diogenes utters this, "with the intention of uttering words *that will be interpreted by Alexander as true* if and only if Diogenes would have him stand from between Diogenes and the Sun" (p. 435, my emphasis). The truth condition of the statement is Diogenes' mental state: this sentence would be true if he does in fact want Alexander to move. As I'll soon show, such an analysis of Diogenes' performing the speech act of requesting Alexander to move from between him and the Sun, in terms of the meaning of the sentence uttered and its truth-conditions seriously mischaracterizes Diogenes' intention. It encourages a focus on this "first meaning", and on speakers' intentions to be interpreted as meaning a certain thing. In many cases, speakers do not operate at this level of description of their speech acts. Interpreters, likewise, don't operate at the level of interpreting first meanings. Such an analysis distorts the picture, by diverting the focus away from where, as I'll argue here, it ought to be in many cases: on speech acts and on the speaker's intention to do something that will count as the performance of a particular speech act.[73]

Here I will present analytic tools that don't distort the picture in this way. I'll first show why Davidson thinks the phenomenon he raises poses a problem, and outline his solution. Then I'll offer an alternative, Embodied Action (that is, speech act) based, approach to analyzing people's linguistic abilities. This approach highlights a cognitive ability that underlies our ability to use language, by supporting all our interactions with one another. This is the ability to adopt what Dennett (1971, 1987, 1991a) calls the intentional stance, and what ethologists[74] call the ability to "mindread"[75]: the ability to explain and predict others' behaviour by attributing goals and other intentional states to them. Michael Polanyi's (1958) account of skill-learning and his analysis of skillful performance lends further support to my claim that speaking language is a skill. I use Polanyi's analysis to support my claim that in exercising this skill in everyday conversation, people principally operate at the level of speech acts and the speaker's purposes in performing the speech act they do, rather than at the level of words and their meanings and how those meanings contribute to the meanings of the sentences uttered. I'll then show how this account smoothly explains the problematic linguistic abilities Davidson brings to our attention, in contrast to the rather *ad hoc* apparatus Davidson uses to explain such abilities. Lastly, I'll also support this speech act approach to analyzing people's linguistic

[73] Moreover, it distorts the picture even further to attempt to extend this type of analysis, so that linguistic interactions are characterized and analyzed in terms of the speaker "asserting a sentence whose (intended) meaning is a proposition the speaker holds to be true" plus some other pragmatic stuff; such as Searle's (1969) notion of appending illocutionary force to the sentence uttered.

[74] E.g. Whiten (1996a, 1996b, 1997, 1998).

[75] Some refer to this ability as possession of a "theory of mind". While I don't want to argue the point here, my focus on speaking a language as an ability, rather than knowing a theory, inclines me away from using this term, in favour of the ability to adopt the intentional stance, or to "mindread" as it is sometimes called.

skills and linguistic interactions by looking at the pessimistic conclusions Davidson himself draws for meaning based accounts of language, in the light of the fact that we *can* deal with "slips" like these. Davidson concludes that a language cannot be the sort of thing many linguists and philosophers have supposed language to be, and that we must give up on such conceptions of what it is to know a language. I argue that the speech act approach I outline here is a viable alternative account of what it is to know a language, one that avoids Davidson's pessimistic conclusions.

4.2 *"Standard" conceptions of language.*

Davidson points out that what he calls "standard" conceptions of the nature of language and of linguistic competence, including views Davidson admits he himself is responsible for (p. 437), are threatened by a consideration of our ability to correctly interpret the sorts of "slips" I just mentioned. Such "standard" views take three principles about the meanings of utterances as the foundations of language-use. Firstly the meaning of an utterance is *systematic,* in that it is given by the meaning of the parts of the sentence uttered and the ways these parts are put together. Secondly, meanings are *shared* by all speakers of that language. Thirdly, meanings are governed by *pre-learned* rules and regularities; we learn the rules and then subsequently apply them to utterances we interpret.

According to these "standard" conceptions of language, before conversing with someone I have an idea of what meanings this person will assign to what expressions, what names they will be familiar with and know the referents of, what phrases they will not be familiar with, and so on. Davidson calls this my *prior* theory of interpretation for this person (p. 442). Possessing this prior theory enables me to interpret this person's utterances, and enables me to make utterances I expect this person is able to interpret.

The linguistic slips Davidson brings to our attention threaten such conceptions of language and linguistic competence, he says, because these slips "introduce expressions not covered by prior learning, or familiar expressions which cannot be interpreted" (p. 437) using the pre-learned rules and regularities for those expressions. An interpreter's prior theory can't serve for correct interpretation when the speaker makes slips of the tongue, malapropisms, or uses ungrammatical sentences or unfamiliar jargon. The phrase "wheel thingummy", for example, is too unspecific to serve any useful function in my prior theory for speech transactions with my roommate, and so would not be a part of that theory. However, when I interpreted my roommate's utterance successfully, I knew that *on this occasion* "wheel thingummy" meant what "the pizza-slicing wheel" would standardly mean. I did this without using a theory of interpretation learned prior to this occasion of interpretation, nor any theory of interpretation I shared with her prior to interpreting the utterance.

To correctly interpret such "slips"—here "correctly" being as the speaker intends and expects them to be interpreted (p. 436, 440)— claims Davidson, I must have temporarily amended my prior theory, since only an amended theory could serve to interpret the speaker's utterance as the speaker expected and intended it to be interpreted. Thus, according to Davidson, I must have constructed what he calls a *passing theory* of interpretation, which is my prior theory, custom tailored to correctly interpret this utterance only, of this person only, on this occasion only (p. 443-4). In the passing theory I constructed to interpret my roommate, then, the phrase "wheel-thingummy" was given all the powers that "pizza-cutting wheel" has in my prior theory for speech transactions with her.

This forces us to revise our idea of what it is to know a language, our idea of what speakers and interpreters must share if they are to be able to understand one another. To Davidson, a speaker and interpreter can't employ a previously possessed and shared theory of what certain expressions or utterances mean. Rather, he says, they must share *the ability to construct a correct, that is a convergent, passing theory of interpretation for speech transactions with one another* (p. 445).

If we are trying to understand people's linguistic abilities, and the support-mechanisms these abilities depend upon, the crucial question here is: "In what does the ability to converge on a successful passing theory consist?" A passing theory, says Davidson, is constructed

> by wit, luck and wisdom, from a private vocabulary and grammar, knowledge of the ways people get their points across, and rules of thumb for figuring out what deviations from the dictionary are most likely (p. 446).

There are no rules to follow in constructing correct passing theories, he says, "no rules in any strict sense, as opposed to rough maxims and methodical generalities" (p. 446). There is also no way of regularizing the process of constructing a successful passing theory, nor is there any chance of teaching someone how to do so. Davidson likens the process of constructing a correct passing theory to the process, in any field, of constructing a new theory to explain new data that can't be explained by an existing theory. Because of this similarity, Davidson concludes that we have "eroded the boundary between knowing a language and knowing our way about in the world generally" (p. 445).

This is not a conclusion Davidson welcomes. Davidson wants to understand our ability to use *language*. His aim is to understand people's specifically linguistic skills and competencies, and to delineate these from other competencies people might have. To Davidson, language-use is all about the interpretation of utterances using a theory of the meaning of the sentence uttered. He notes at one point that what he calls "first meaning"—the meaning a person intends their utterance to be interpreted as having— is not limited to language; according to everything he's said so far, "first meaning" applies to any sign or signal with an intended interpretation. So, he asks, "what should be added [to what he's already said] if we want to restrict first meaning to *linguistic* meaning?" (p. 436, my emphasis). His aim is to make just such a restriction. Davidson aims to account for linguistic meaning, not meaning in general. One of his principal questions is: to what extent should the various competencies that interpreters have be considered linguistic? (p. 437). Part of the burden of his paper, says Davidson, is "that there is much that they [that is, interpreters] can do that ought not to count as part of their basic *linguistic* competence" (p. 437, his emphasis).

I disagree fundamentally with this approach. I welcome the conclusion that we've eroded the boundary between knowing a language and knowing our way about in the world generally (although I disagree with Davidson's reasons for drawing this conclusion). In contrast to Davidson, I am severely disinclined to draw any sort of sharp distinction between people's linguistic abilities and their non-linguistic abilities. I believe that many competencies involved in understanding other people's utterances are not specifically *linguistic* competencies, but competencies that we apply to a broad range of social interactions and practices. One of my central points is that much of the skills and knowledge that enable us to speak with and understand one another are not *separably* linguistic in nature, as Davidson supposes, but are part of our general abilities to successfully participate with one another in shared practices. Separating linguistic interactions and linguistic skills from the rest of our human interactions and skills, I believe, is a symptom of the (misguided) desire for a simple, general theory of language.[76] This desire, combined with analytic tools designed for studying sentences and their meanings and/or truth-conditions, gives rise to an over-simplified theory, that is over-generalized when used to analyze all linguistic phenomena. Using such analytic tools to explain people's linguistic competencies is a mistake. It encourages a separation of people's linguistic skills from the more generally applicable skills and knowledge that support our ability to perform and interpret speech acts while participating together in the shared practices that comprise our forms of life.

This approach, in explaining what people learn when they learn a language, also encourages a focus on *generally-applicable context-independent* theories of meaning for particular words or sentences (or norms for what expressions ought to be used to mean). This means that, to account for a particular aberrant use in a particular situation such as the kinds of utterances Davidson highlights, the general theories need to be modified in rather ad hoc ways. This suggests that people are continually employing these ad hoc modifications to general theories, and that learning a language is learning the skill of constructing such ad hoc modifications to the general theory of meaning one learns.

A more productive approach, as I'll soon argue, is to look at the particular use being made in each particular instance, and at the practice within which this use "has its life", and not aim at a general theory of meaning for particular expressions. There may be generally applicable sets of norms in use in people's linguistic exchanges, but these are not the norms of any general theory of meaning. These generally applicable rules are the norms governing practices and how one ought to act when participating in them. Particular uses of expressions, even aberrant ones, get their "life" –their role in human lives – by virtue of the practice within which they are being used. Thus general norms are

[76] I call this desire for generality misguided, because here "generality" is used as synonymous with "context invariant". This desire for as little context-dependence as possible underlies the desire for generality. But on the account I present here all uses of a particular expression or sentence are context-dependent; a particular expression can be used to perform many diverse linguistic tasks. And the purpose an expression can be put to, can vary tremendously; especially when we don't ignore metaphorical uses of language, as many theories of meaning do. There are certainly purposes that each expression is well-suited for performing, but given a particular context and history, almost any expression can be put to the service of a rather unusual purpose, and the audience can still appreciate the purpose it is being put to in this instance.

employed in shared social practices without the need for frequent ad hoc modifications.

4.3 *A Speech-Act Approach to People's Linguistic Abilities*

As I've been saying, I believe we should focus on language-use not as uttering meaningful expressions, nor even as uttering meaningful expressions with illocutionary force but as a species of performing purposeful actions within shared social practices, using linguistic tools to perform these actions effectively. Grice's insightful account of the nature of linguistic exchanges helps show what I mean here[77] (although I'll be modifying the detail and the focus of Grice's account). To mean something by an utterance is not, as Searle (1969, p. 43 ff.) interprets Grice, the same as uttering a sentence and meaning it. To mean something is to utter something as part of an interaction with someone else, and to intend to produce a particular kind of response in that person as part of your interaction. Grice has a particular idea of what the means of producing that response is. To Grice, a speaker *S* "means something" by uttering *x*, if and only if *S* intends:

(1) that *x* have certain features, *f*;

(2) that a certain audience *A* recognize (think) that x has features *f*;

(3) that *A* infer at least in part from the fact that *x* is *f* that *S* uttered *x* intending,

(4) that *S*'s utterance of *x* produce a certain response *r* in *A*;

(5) that *A*'s recognition of *S*'s intention (4) shall function as at least part of *A*'s reason for producing response *r*.[78]

I have much sympathy with Grice's focus on the interactive nature of linguistic interactions; especially his emphasis on the speaker's intentions, and in particular on the speaker's intention to produce a certain response in the audience by means of getting the audience to recognize the intention to produce that response. I want to modify this account in two ways, however: one minor and one major. None of these modifications are completely absent from Grice's

[77] See H. P. Grice (1989), Ch 5 "Utterer's meaning and Intentions" pp. 86-116, and Ch 14 "Meaning." pp. 213-223. All page references to Grice are to this book unless otherwise stated.

[78] This is Stephen Schiffer's (1972) formulation, with recommendations from Grice, of Grice's account of S-meaning.

account. They all however, are under-emphasized and need to be drawn out of Grice's explanations and made explicit. As well as modifying Grice's account, I also want to shift the focus, away from the focus on the speaker "meaning something" –and thus away from the focus on what exactly it was that the speaker meant– and instead on the speaker achieving the intentions with which the speaker acts.

The minor modification is to (1). It's not just the words I utter that have the features you need to recognize when I speak to you. Rather it's my act of uttering these words, to you, in this situation, with this history and mutual knowledge between us, in this shared form of life, that has the relevant features. It's these features of the physical, social and historical context that help the audience identify the practice that the speaker's utterance either initiates or is situated within.

The major modification I want to make is as follows: Point (3) states that I intend that you infer my intention to get you to produce response *r* by way of recognizing the features *f* my utterance has. Grice says that the speaker intends (among other things) that the audience "think of *f* as correlated in way *c* with the type to which *r* belongs" (pp. 103, 104). It's these correlations that I want to focus upon. It needs to be emphasized that there are usually two intermediate intentions that I also intend you to recognize. I intend that you recognize a certain intention, on the basis of features *f*. I also intend that your recognition of this intention bring to mind the way my action of making this utterance is "correlated in way *c*" with the response I intend to prompt you to make.

The intermediate feature is this: you must recognize what illocutionary act I intend to perform. This modification is made to the analysis to cover the fact that I can perform an illocutionary act like asking you to close the door by uttering all manner of phrases. I could utter "please close the door", or I could use a more obscure colloquialism, like "Put the wood in the hole." If you habitually leave the door open, I might even just yell "Hey!", intending that uttering this, in this situation, with that history between us, will be sufficient to make you recognize that asking you to close the door (while also reprimanding you for again having to be asked). I could even ask you to do this using non-linguistic means, such as pointing at the door and glowering disapprovingly, or simply throwing your things out the door until you close it. Some of these actions will make my intention to ask you to close the door more recognizable to you than others.

Furthermore, that I am performing this speech act, is the basis of your recognition that I'm engaging in the practice of asking you to do something. Your recognition that I am invoking the practice of asking you to do something and the norms that govern this practice structure our interaction. These norms (correlations *c* to Grice) dictate several ways it would be appropriate for you to respond (refusing, promising to do it soon, complying, etc.). They also help you recognise that people engage in this practice of asking you to do something as a means of getting you to do as they ask. Thus you can recognise which of the appropriate responses is the one I want you to produce: I do this *because* I am trying to get you to close the door.

So, in addition to modifying some of the above points I want to add the following points. *S* also intends:

(2.1) that *S*'s act of uttering *x* (with features *f*) count for *A* as the performance of illocutionary act *i*.

(2.2) that *S*'s performing illocutionary act *i* (by performing the act of uttering *x* with features *f*), be recognized by *A* as a legitimate move within practice *p*.

Point (4) will also need to be extensively modified to reflect the way that the norms of the practice are also the basis of A's recognition of the type of response that *S* intends that *A* produce. Furthermore, point (3) needs to incorporate the fact that *S* intends that *A* recognise *S*'s *reason* for performing this action (uttering *x* with features *f*) is *because* *S* intended (4). *S*'s *reason* for doing all this was to get *A* to respond in this way. Thus:

(3) that *A* recognize (or infer)[79], at least in part because of the fact that *S*'s act of uttering x has features *f*, that the *reason* *S* uttered *x* (with features *f*) is *because* S intends:

(4) that *A*'s recognition of intentions (2.1) and (2.2) prompt *A* to produce a certain response *r*, (the kind of response that –within practice *p*– is appropriate and that performing *i* is a conventional means towards).

[79] This could be cast as a recognition or an inference. Sometimes the intended response needs to be figured out, and sometimes, when the norms of practice p are familiar and tacit enough (A is a connoisseur of the practice, in Polanyi's terms –see section 5 of this chapter) that "this is just what we do", the intended response may just be obvious. From now on I'll stick to "recognize". When I use "recognize" however, I do not mean to imply that error is impossible. Here I use "recognize" in the sense that I can recognize an x as y when x is not in fact y.

Recognizing which illocutionary act the speaker intends their utterance to constitute the performance of, is essential for the success of any illocutionary act. It's one of the more important of Austin's felicity conditions on any illocutionary act: illocutionary "uptake" (pp. 22, 138) must be secured. If the speaker can't make the kind of speech act they intend to perform recognizable, then the speech act is infelicitous; by Austin's classification, it will be "void" (p. 22). And the audience's recognizing the practice that this speech act is intended to be a move within is essential to the speaker's awareness that this "uptake" is secured, since the usual way of knowing whether it has been secured is that the audience responds with a legitimate response, according to the norms of that practice. For instance, when I try to apologize to a friend for not meeting her as we arranged, she must recognize that an apology (and not, for example, an excuse or a taunt) is being offered. She must also recognize that I'm apologizing for not meeting her, not apologizing for arranging the meeting. If what I am doing is not recognizable to her as an apology for not meeting her, then I haven't successfully apologized for not meeting her. Furthermore it is by her engaging with me in the practice initiated by my offering an apology, by acknowledging my apology (perhaps also forgiving me) that is the criterion of my having secured "uptake". Her acknowledgement is confirmation to me that she counts what I did as an apology. This condition of securing illocutionary uptake applies similarly to invitations to meet for coffee, promises that I will be there next time, assertions that I failed to show up because I had a flat tyre, and all other speech acts that initiate a practice which the audience responds to by participating in that practice as well.

Additionally, the connections *c* that Grice speaks of (e.g. p. 103, 104) between the features *f* of the utterance and the intended response *r*, are often connections between types of speech act and conventional responses to that type of speech act; connections secured by the norms governing the practice these actions are situated within. It's very often the case that the ground of your recognition of the response I intend to produce in you is your recognizing that my utterance is the performance of a particular illocutionary act and that this act is a generally accepted way of initiating a particular practice. The action I perform and intend you to respond to, is tied through custom and convention –through the norms governing the practices we learn how to participate in as we become socialized members of our forms of life– to the types of action I expect you to perform in response. The action I intend you to perform in response is the *reason* that I perform the action. And this reason is recognizable to you because of these norms of means to ends that the practice institutes.

Interpreters' attention to a speaker's *reasons* for the moves they make within practices is also apparent in Grice's (1975, 1989, ch. 2) mechanism of *conversational implicatures,* based in the cooperative principle. The principle's attendant maxims of quantity, quality relation and manner, relate not to the speaker's reason for performing the move they do (although they can help with interpreting that, too), but principally to their reason for *the way* they make that linguistic move. They give their act of making an utterance features *f* for a reason; often this reason is not directly related to the type of move they make within the practice, but to the way they make that move. For instance, people often *intentionally* depart from these norms, and the reason for doing so can often be obvious to an interpreter. The classic example is the professor who, when asked to write a reference for student X, writs "X has nice handwriting and was always punctual". The norm would be to write about the student's academic abilities. The person receiving the letter knows this norm, and believes that the professor also knows it and is deliberately flouting it, by giving less information than would be *norm*al. The recipient can easily recognize the professor's reason for deviating from the norm: the professor thinks the student has no academic abilities worth writing about. Alternatively these norms can help identify implicatures because, for example, although a speech act can seem irrelevant, by being set within the practice its presumed relevance carries an implicature. For instance, Grice (1989, p. 32) gives the following example:

> A: Smith doesn't seem to have a girlfriend these days
> B: He has been paying a lot of visits to New York lately.

Because B makes this reply to A's remark, A can infer that B intends his reply to be relevant, and so recognizes that B's reason for making this remark is to implicate that Smith has or may have a girlfriend in New York. These general norms governing the process of making a speech act as a move within a norm-governed practice can also assist interpreters in deducing the speaker's reasons for doing what they did in the way that they did it (uttering *x* with features *f*).

A good example of the way a conventional practice can structure the relationship between a speech act and appropriate responses to it, is the practice of asking someone to do something. Winograd and Flores (1986, p. 65) give this "map" (fig 1) of the *linguistic*[80] moves that are opened up by person A requesting person B to do something. At each stage of the practice of asking someone to do something, only a small set of possible linguistic moves are "open". For instance, after A's request (at stage 2), B can promise to do it later, B can counter with an amendment to the act requested, B can reject the request, or A can withdraw the request. Other possibilities include B's questioning the intelligibility of the request ("I didn't hear you"), or questioning the authority of A ("you can't order me to do that"), or B's simply performing the act requested. Knowing which possibilities are "open" at each stage, we expect that a subsequent speech act will be performed with the intention that this action constitutes one of the "open" moves. Each of the participants expect the other's utterances to either fit this pattern by being the performance of one of the "open" moves in the interaction, or to change the subject. However in some circumstances, changing the subject could not be an "open" move. The "dance" isn't over until it reaches one of the "terminal" points (the circles with the thicker outlines: 5, 7, 8 or 9).

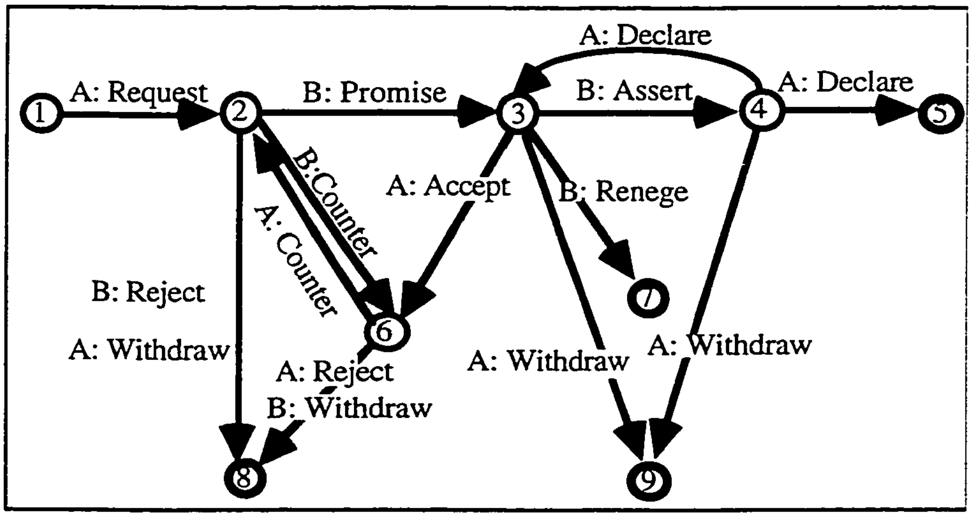

Fig. 1 (Adapted from Winograd and Flores (1986), p65.)

Most interactions, including conversations, take the form of mutual leading-and-following "dances"[81] like this. The practice we're participating in

[80] Note that they do not include the non-linguistic moves, such as B's performing the act requested. Since an appropriate response to the speech act of asking someone to do something is to just do it (and not even need to declare it done). It would be appropriate to include these moves in the diagram as well.

[81] For more on this metaphor of conversational "dances", see Winograd and Flores, Op Cit., pp. 64-5.

(just as if we were waltzing, two stepping or tangoing) structures the kinds of moves that are appropriate and the kinds of responses to our partner's moves that one should make if one is to continue to participate. I do something, intending that you recognize what I did and that you to "follow my lead" by responding to what I do. And I, in turn, follow your lead, responding to what you do. What we each do conforms to the norms governing the practice; norms that specify appropriate responses to others' actions.

The norms also specify relations of ends to means, such that the response I hope my conversational partner will make is often also apparent as the *reason* that I did what I did. For example, a customary reason for asking someone to do something is to get them to do it. A customary reason for informing someone of something is to get them to believe it. A customary reason for asking someone a question is to get them to answer it. Speakers depend on their audience's familiarity with these practices and their norms to frame their utterance as motivated by a particular recognizable reason. In performing a certain illocutionary act, then, I expect that you will share with me enough of a background of cultural and linguistic practices (we share a "form of life" in Wittgensteinian terms) that you will be able to recognize (or infer) that the reason I did what I did is because I intend you to respond in a certain way. And you can recognize how I'm trying to get you to respond, because you similarly expect that I share with you this form of life, which includes practices and their norms of linguistic means to ends. Because we both know how to participate appropriately in these practices you appropriately and justifiably expect that I rely on your attributing to me ends that follow, according to the practice's norms, from the means I just used.

I talk about *knowing how* to participate in these practices deliberately. As I argued in the last chapter, many of the norms that structure our practices can be quite tacit. We could make them explicit (as Winograd and Flores did for the practice of asking someone to do something). But we follow them and they structure our interactions and guide our expectations about the next move in the interaction, even if they do so tacitly.

I should now re-state the full set of points about the speaker's intentions when a speaker performs a speech act, with all of these alterations and additions. A speaker, *S* performs a "meaningful" speech act in making an utterance, if and only if, *S* intends:

(1) that *S* 's act of uttering *x* have certain features, *f*;

(2) that a certain audience *A* recognize that *S*'s act of uttering x has features *f;*

(3) that *S*'s act of uttering *x* (with features *f*) constitute for *A* the performance of illocutionary act *i;*

(4) that A recognize that *S*'s performing illocutionary act *i* is a legitimate move within practice *p;*

(5) that *A* recognize, at least in part because of the fact that *S*'s act of uttering x has features *f*, that the *reason S* uttered *x* is *because* S intends:

(6) that *A*'s recognition of intentions (3) and (4) prompt *A* to produce a certain response *r*, (the kind of response that –within practice *p*– is appropriate and that performing *i* is a conventional means towards);

(7) that *A*'s recognition of *S*'s intention (6) shall function as at least part of *A*'s reason for producing response *r*.

Let's flesh this all out by putting our earlier example into this schema. As you stand there in the doorway taking your boots off as the cold air leaks into the house, I utter "Will you please close the door" to you, intending:

(1) that my act of uttering this phrase has certain features: that these particular words are uttered, by me, to you, in this particular physical context, with that particular history of interactions between us, in this particular shared "form of life";

(2) that you recognize that my act of uttering this has those features;

(3) that (on account of recognizing those features) you count my utterance as the performance of the speech act of asking you to close the door;

(4) that you recognise the speech act of asking you to close the door is a legitimate move within the practice of asking you to do something.

(5) that you recognize that the *reason* I performed this speech act is *because* I intend:

(6) that your recognizing that I'm asking you to close the door (initiating the practice of asking you to do something) as a means of prompting you to close the door;

(7) that your recognizing that I'm trying to get you to close the door function, as part of your reason for closing the door.

Here I intend (2) that you recognize that I've given my act of uttering words certain features. I also intend (3) that you recognize that an action with these features is intended to count as the performance of a particular speech act, and as a move in a certain practice that I assume we both know how to participate in felicitously. I further intend (5) that you recognize (or infer) *why* I am performing that speech act. I intend (6) that you recognize that my reason for asking you to close the door is that it's a conventional means of getting you to close the door. In addition, I intend (7) that your recognizing my further aim of getting you to close the door will be part of your reason for responding as I intended you to.

These features of the act of making, and interpreting, an utterance illuminates the difference between performing an illocutionary act, and achieving a perlocutionary effect by doing so. Searle (1969, p. 44 ff.) claims that Grice defines "meaning something" by an utterance in terms of performing a perlocutionary act (producing a response in the audience). But "saying something and meaning it is a matter of intending to perform an illocutionary, and not necessarily a perlocutionary, act" (p. 44). Searle's objection elides the point that one cannot simply perform a perlocutionary act. To perform a perlocutionary act is also, and *eo ipso*, to perform an illocutionary act; the illocutionary act is the means of achieving a perlocutionary effect. My point is that in order to produce a response in the audience, the speaker intends that the audience recognise what illocutionary act the speaker intends to perform and which practice that this illocutionary act is a move within. The speaker also intends that the audience produce one of the responses that the norms of that practice stipulate as appropriate responses to the performance of that speech act. The audience's recognition of the illocutionary act performed and the practice it is a move within, therefore, is the means of producing the intended perlocutionary effect.

In making this objection, Searle concentrates on the *words the speaker utters* and what they mean, and loses sight of the speaker's *act of uttering* it and the practice this is a move within. This is apparent in his supposed counter-example to Grice's account of a American soldier in WWII captured by Italians, who speaks a line of German poetry learned in high-school, hoping that by speaking German he can get the Italians to believe that he's a German officer and so set him free. Searle thinks of the speaker's primary intention as an intention to *mean* a certain thing (that I am a German officer), rather than an intention to get them to attribute to him an intention to engage in the practice of inform them that he is a German officer. Searle objects that the American can't intend to mean "I am a German officer," when he knows that the phrase he speaks really means "Do you know the land where the lemon-trees bloom?" This objection misses the point. What the phrase spoken *means*, and even what it is intended to mean, is almost completely irrelevant.[82] This line of German is uttered with the intention that the Italians recognize that he is speaking German, and speaking it in the way they expect an officer might speak. The speaker intends that (partly because of the context etc.) they attribute to him the intention to engage in the practice of informing them that he is a German officer. He also intends that they also engage in the practice he initiates and respond appropriately. He intends that they believe that he is the right kind of person to be legitimately engaging in that practice, and attribute to him the reason such a person would have for speaking to them in the way that he did, and respond appropriately to that intention that is the typical reason for performing the speech act of informing them that he is a German officer, by treating him as an ally and not a prisoner.

One of the problems for accounts of language like Grice's, is this assumption that in every case of linguistic utterances where a speaker "meant something", we can extract and explicitly state what it was that the speaker meant.[83] Many of Grice's examples and reformulations of his analysis (see, for example Grice 1989, chapter 5), are to make room for this idea that if the speaker meant something, we must be able to identify what it was that the speaker meant. Often, however, (this is the case with Searle's example) to attempt to identify what the speaker meant is largely an attempt to identify something that played little role in the actual interaction. What is important in understanding the mechanics of the interaction is not what the speaker meant, but what speech act the speaker intended his action to count as for the audience and what practice the speaker intended to engage the audience's participation in. The American did not intend the Italians to respond to the meaning of his words; he assumed that the Italians would *not* attach any particular meaning to the words he used. Rather he intended them to respond to the fact that he is speaking German (in an authoritative tone, etc.), and to believe that he is trying to perform a speech act, where at least part of the reason for doing so is to inform them that he is a German officer. The primary intention of the speaker is to get the Italians to respond in ways the practice's norms stipulate as appropriate, to the reason he

[82] Grice also points out the irrelevance of what the line spoken itself means (Op Cit. p. 102) and that it's the act of making this utterance that is supposed to provoke the intended response.

[83] John Searle is a good example of someone who takes this to heart. His Principle of Expressability (Op Cit. p. 19) stipulates that whatever can be meant can be stated explicitly.

performed this illocutionary act, not to any imputed meaning of the sentence uttered.

I'm not advocating this model as adequate for analyzing *all* linguistic phenomena. I'm sure that counter-examples can be constructed that will test its adequacy. I am suggesting, however, that this model is useful for analyzing *many* linguistic phenomena; especially when our focus is on people's linguistic abilities. I'm also suggesting that the centrality of the ability of both speaker and audience to attribute intentions and beliefs to one another (by virtue of the norm-governed practice their actions are set within) is no accident. These abilities and the practices within which they are exercised *are* centrally important, and that importance needs to be highlighted.

This analysis also works for many non-linguistic actions as well as linguistic ones, where the agent's purpose in performing the action is to produce a response in another by means of making recognizable the agent's intentions to engaging the other in a certain practice. For instance, imagine that someone made eye-contact with you in the street, and tapped the back of their (bare) wrist, while looking expectantly at you. Assuming a shared cultural background that includes the practice of wearing a watch on the wrist, it's pretty easy to recognize that this person intends their action to be recognized by you as asking you to tell them the time of day, Furthermore, they expect that you would recognize that this is done in order to engage with you in the practice of asking and telling the time, and this as a means of getting you to tell them the time. Likewise, responding with a shrug of the shoulders and displaying bare wrists could function, within the practice the other initiated, to informing them that you don't know what time it is (in order to get them to believe that would tell them the time if you knew what it was, but that you are unable to do this).

We need to keep in mind these similarities between linguistic and non-linguistic interactions, and between the ways they both can be analyzed within shared social practices. As Grice notes, "...surely to show that the criteria for judging linguistic intentions are very like the criteria for judging non-linguistic intentions is to show that linguistic intentions are very like non-linguistic intentions."[84] It's because of these similarities that I resist the push towards isolating and analyzing people's especially *linguistic* skills and competencies, as Davidson and others recommend. A great deal of our linguistic skills and

[84] H. P. Grice, "Meaning.", Op Cit. p. 223.

competencies are, at bottom, more general skills and competencies for participating in practices within which performing recognizable speech acts is are one type of move among many.

4.4 The Development of Language and Mindreading in Infants

On the account I just presented, the ability to "mindread" others intentional states, as it's often referred to in ethology and developmental psychology, by observing their actions, is one of the central skills and competencies that people's ability to use language depends upon. This is practically an intuitive ability for us. As Terrance Deacon (1997) rightly points out (p. 416), some children are mathematical savants, and can intuitively "see" answers to very abstract math problems that most others can only deduce through laborious calculation. In a similar way, all children nowadays (with a few exceptions I'll discuss presently) are language savants; we are all born with the genetic predispositions to be exceptionally gifted at learning to use language. A large part of this is due to human children being similarly gifted at learning to mindread. Language abilities come with mindreading abilities; it's a package deal. I'm going to show this with two brief surveys. In the first I'll look at the ways mindreading skills develop in normal healthy children. In the next section I'll contrast this normal development with a survey of autism. In autistic people, many of our language-specializations (such as grammatical abilities, memory for words and their referents) are present, but where the ability to mindread does not develop to support these more specialized linguistic abilities. The aberrant linguistic and social interactions demonstrated by autistic people further highlight the ways in which mindreading skills support the social and linguistic practices that, according to the account just presented, depend upon their exercise.

In normal, healthy children, the ability to mindread begins to develop very early on. Developmental psychologists often refer to the development of a "theory of mind"[85] to refer to the way this ability develops in children. Much

[85] From my orientation in previous chapters it should be clear why I prefer a characterization in terms of an innately disposed set of rather tacit mindreading abilities, rather than an innately developing theory of mind. I use the term reluctantly here, simply because it's the term most often used in the literature. There is a debate in this literature, about whether this theory of mind is best characterized as a Theory theory or a simulation theory; as the ability to apply a theory about others' intentional states, or the ability to use one's own intentional states to run a simulation, of the others' intentional states. ("If I were in their position, I would think/want/intend/etc.). Although I have inclinations towards the latter, I do not think it leans far enough

"theory of mind" development research focusses on the "Sally-Anne" test for the ability to attribute false beliefs to others, which normal children pass at around three and a half to four years old. Before this age, children are notoriously unable to attribute beliefs to others that are different from their own beliefs. Carol Feldman, however, refers to this as a "Graduation Day" test (Bruner and Feldman 1993, p. 269). Concentrating on the "Sally-Anne" test (as many do), she says, obscures the contribution to a child's mindreading skills of the preceding three or four years of development. Children much younger than three and a half already demonstrate great sensitivity to other peoples' perspectives, their intentions and goals, and their states of knowledge and awareness.

We can trace this development back to certain innate abilities, apparently present at birth. Meltzoff and Moore's (1977, 1983) experiments, for instance, show that infants "imitate from birth". As soon as they are born (in one study the average age of subjects was 32 hours, the youngest was 42 *minutes* old), infants are able to imitate adult's facial gestures (tongue-poking, mouth-opening and lip-protrusion). At twelve to twenty-one days old, the infants tested could even do so from memory when disabled from imitating immediately by being given a pacifier to suck on (1993, p. 342). That they can do this in spite of the fact that they cannot see how their own facial gestures look, suggests a very early awareness of a connection between how others look and the way the infants feel themselves to be. Between nine and twelve months old, infants begin to develop an understanding of others as experiencing events, and an understanding of events and objects being the subject of someone's attention (Wellman 1993). Mutual-imitation-games resembling gestural turn-taking "dialogues" between parents and young infants, also assist in the development of an awareness of other's perspectives. Perhaps partly because of such games, at fourteen months old infants demonstrated that they could appreciate when an adult is imitating their actions (Meltzoff and Gopnik 1993), indicating an awareness of how they appear to others. Meltzoff (1995) has also shown that eighteen month-old infants can apparently discriminate between the goal-driven actions of an agent and not goal-driven but practically identical movements of a mechanical device.

towards the largely tacit and practice-situated set of abilities I am talking about. But I do not know enough about the debate to declare much more than this prima facie inclination.

Particularly relevant here are infants' developing abilities to understand other agents' actions as goal-directed and to recognize what others' intentions are, and also the ability to make their own intentions recognizable to others. Language-use, as I've portrayed it, relies on both conversants in a linguistic interaction possessing these abilities. Not coincidentally, these abilities appear to develop at about the same time that linguistic skills begin to emerge. For instance, some have documented pre-linguistic infants (nine to twelve months) engaging in purposive "communication" (Bretherton, McNew and Beeghly-Smith 1981), in that they make attempts to direct others towards desired objects, by way of pointing and gesturing, by gesturing with eye-contact alternating between the person they want to get the object for them and the object itself, and by continuing to amend their gesturing until the goal has been attained.

Vocal and gestural interactions that rely on the ability to attribute wants or goals to others appear to develop shortly after this, at about twelve to thirteen months old; for instance, Masur (1983) reports that infants are able to interpret pointing gestures by adults not simply as relating to an object, but as a request to be given the object. Thus infants can play games where the object requested by an adult is repeatedly offered then snatched away by the infant at the last moment, while watching the adult and smiling, apparently taking delight in thwarting the other's desires (Reddy 1991). Infants at eighteen months old have been shown to demonstrate an awareness of what someone intends to do by imitating the actions an adult was trying unsuccessfully to perform (Meltzoff 1995). They were able to attribute intentions to others, and imitate the goal-action even when they never saw the adults attain the intended goals.

As I said earlier, the "graduating" achievement, in the development of mindreading abilities is taken to be the child's ability to pass the false belief test. Recognizing that someone' else's beliefs are different from their own (when they recognise others' beliefs *as beliefs*) occurs between three and a half to four years old.

Thus infants develop certain mindreading abilities very early on; sophisticated forms begin to emerge at about nine to twelve months, and flourish around eighteen months. Interestingly, their abilities to make their purposes recognizable to others, and to attribute purposes to others, develop from twelve to eighteen months old, at about the same time that most infants also begin participating in linguistic interactions; they begin to emerge at about nine to twelve months, but flourish around eighteen months.

These innate abilities, or their genetically programmed development at the appropriate time is evidence of a long history of selection for the ability to mindread within human cultures. I'll come back to this point in Section 6.2.

4.5 Autism: a Failure to Develop "Mindreading" Skills

That language and mindreading develop at about the same time is hardly surprising, if our participation in linguistic interactions depends upon the ability to attribute intentional states, especially goals and intentions, to others. We can get some idea of just how essential the ability to mindread is to both our social and our linguistic interactions by looking at autism, where this ability fails to develop.[86] As Austin (1961, p. 128) counsels us in 'A Plea for Excuses', "... as so often, the abnormal will throw light on the normal, will help us penetrate the blinding veil of ease and obviousness that hides the mechanisms of the natural successful act." Looking at the research on autisim, affords a glimpse behind "the blinding veil of ease and obviousness" of the dependence of normal our "natural successful" interactions on our mindreading abilities.

Autistic individuals are apparently severely impaired in their ability to interact socially with others, and especially in their ability to interact linguistically with others, because they cannot appreciate other people's intentional states; they're unable to "mindread". Normal healthy people take it for granted that some of the things they encounter are agents; agents with obvious goals and purposes, beliefs and desires. This "natural psychology" helps us make sense of the behaviour of agents. Other peoples' beliefs and goals are almost completely invisible to autistic people, however. Imagine what it might be like to be unable to recognize the beliefs and goals that drive other people's behaviour. Alison Gopnik describes the experience of a person who suffers from autism, like this:

> "Around me bags of skin are draped over chairs and stuffed into pieces of cloth, they shift and protrude in unexpected ways... Two dark spots near the top of them swivel restlessly back and forth. A hole beneath them fills with food and from it

[86] The failure to develop a "theory of mind" is one of the more widely accepted explanations for autism. See, for instance, Carruthers (1996), Leslie (1991) and many of the contributions to Baron-Cohen (1993), which is devoted to debating the theory that autism is characterized by a deficit in the development of a "theory of mind". I'm not going to defend the theory of mind deficit explanation here. I take it that the above authors have done an adequate job of that, and defer to their arguments . I'm most interested in the picture we get if we assume this to be a good explanation, whether or not there are also "deeper" explanations for this and other peculiarities to autism.

> comes a stream of noises. Imagine that the noisy skin-bags suddenly moved towards you, and their noises grew loud and you had no idea why, no way of explaining them, or predicting what they would do next" (Quoted in Whiten 1998, p. 5.)

Without this ability to attribute beliefs and especially goals to other people, we would be unable to distinguish agents from non-agents; agents would be things which "shift and protrude in unexpected ways." An autistic person might view other people in much the same way that we might view a robot which moved in apparently random ways and which defied any attribution of purpose, function, or goal. It would be impossible to interact, either cooperatively or to competitively, with such a device.

It appears that it is only by detecting cause-and-effect regularities in people's behaviour that autistic people are able to learn to interact with others, and some eventually learn to use language, albeit in rather odd ways. Bruner and Feldman (1993) refer to this as the Sigman-Feldman 'computational surmise': "that high-functioning autistic children convert the personal world of intention-regulated social experience into an impersonal world of causally-driven events" (p. 288). Thus, they must use their general reasoning abilities to compensate for their lack of mindreading abilities (Leslie and Roth 1993, p. 103); they must *figure out* what other people think and believe and want, through concentrated calculation of the solutions to difficult cause-and-effect problems. Recall the analogy I made earlier: normal people are mindreading-savants in comparison with the rare autistic person, just as some rare people are mathematical savants in comparison with the majority of people. Just as some people can easily "just see" solutions to math problems that the rest of us can only figure out through laborious calculation, some autistic people can only figure out, through laborious calculation, the beliefs and goals of others that most people can "just see" in their actions.

By this tactic, some autistic people can eventually learn to "adequately" function interpersonally, and a small number even learn to use language. However, only the relatively intelligent ones do so with much fluency, and even they very often use language rather oddly. Autistic language-users have, as Bruner and Feldman (1993, p. 288) remark, "a strangeness of talking and manner, that gives normal interlocutors a disturbing sense that they are interacting with someone who is, as it were, outside the culture." It's this same "strangeness" seen from the other side of the interaction, that leads Temple

Grandin, a highly intelligent autistic woman, to describe her social interactions as feeling like she's "an anthropologist on Mars", giving Oliver Sacks (1996) the title for his book.

With a great amount of mental effort and experience, some autistic people can even deal with language which is used to refer to people's emotional and mental states. For instance, in one study, high-functioning autistic people responded adequately to questions about affective states of characters in a videotape, but "the rate, latency and manner of approaching the questions were more like what one expects in a person trying to do a hard arithmetic computation in their head, than what one expects in a person reporting about how others might be feeling." (Bruner and Feldman 1993, p. 278).

The oddities of their language-use can tell us a lot about the role of mindreading in language-use. Carol Feldman concludes from her studies into which aspects of language and linguistic competence are affected by autism, that many of the deficits in their language "suggest a common failure in being able to encode the arguments of action into a structure: an agent pursuing a goal in some setting, etc. (Bruner and Feldman 1993, p. 274). This lacuna in the ability to attribute reasons or goals to others is seriously debilitating to a language-user. Normal people are able to explain others' actions by reference to the agent's reason for performing them, and are able to respond to people's utterances with an awareness of their reason for uttering them. These abilities are largely absent in autistic people. The way that such an impairment would disable people from normal linguistic interaction could be expected to fit certain patterns; the patterns that it appears autistic language-users do display.[87]

[87] Many of the peculiarities of autistic people's language-use are detailed by Tager-Flusberg (1993), Loveland and Tunali (1993, especially pp. 251-260) and by Bruner and Feldman (1993). I've relied on these accounts, and on the sources they cite for most of the details I present here. However, one of the difficulties for my purposes, is that most of the research on autistic people's use of language concentrates on language as communicating information, and not as much on the interactive nature of language-use that I'm trying to highlight. For instance, Tager-Flusberg (1993, p. 153) concludes:

> The specialized and multifaceted forms and functions of human language appear to have been designed for communicating with other people about mental states. One of the primary functions of language, to serve as a major source of knowledge, is impaired in autistic children even in the prelinguistic period. It is this impairment which links deficits in joint attention, later problems with communication, and the understanding of belief.

Because of this prevalent focus on communication (especially the communication of information about mental states, and thus the transfer of knowledge), and downplaying the interactive aspects of language-use that I believe to be important, in a few instances

The major peculiarities of autistic people's language-use are with so-called "pragmatic" aspects of language (Tager-Flusberg 1993, p. 143-4); autistic people have trouble with taking others' knowledge into account in structuring their utterances, with comprehending a speaker's intentions, with identifying different speech acts, and with speaker-listener relations. For instance, because autistic people appear to be insensitive to the knowledge or ignorance of others, their explanations can often be difficult to interpret and relatively uninformative. Examples of autistic people attempting to explain how to play a board-game were woefully uninformative (Loveland and Tunali 1993), because of the subject's apparent inability to recognize the differences between what they themselves know and what others don't yet know. Their instructions appear to be more like descriptions, and descriptions for someone who already knows what is being described; there is little effort to convey the information that someone unfamiliar with what is being "explained" would need to know, and a tendency to refer to things or persons unknown to the listener (Loveland and Tunali 1993, p. 259).

Autistic people are also prone to interpret others' utterances quite literally (Tager-Flusberg 1993); for instance, replying to an utterance used to make an indirect request, "Can you do such and such?" with something to the effect of "Yes, I can." In a similar vein, there are two possible ways a person could interpret someone uttering to them "It's raining outside."[88] The way most normal people interpret this, is to recognize *why* the speaker utters this to you: typically because they intend that you recognize that they intend to *inform* you that it's raining outside, probably because they intend you to believe that it's raining outside. Your coming to believe that it's raining outside thus depends on your estimate of the veracity of what they're informing you of. (Would they inform you of something they know to be false? Could they be mistaken?) It appears that autistic people are less able to comprehend language in this way.

It would seem likely that much of the time autistic language-users simply form literal associations between utterances and facts: deducing that if they hear someone say "It's raining outside," then it must really be raining outside, or perhaps it is very likely to be so. They appear to simply note cause-and-effect relations between states of affairs and utterances —the fact that it's raining

the reasons I offer to explain the patterns researchers observe in autistic language-use differ in focus from the reasons offered by the experimenters themselves.

[88] This example is based on a distinction made by Jonathan Bennett(1976), 193-196.

causes someone to utter this— without appreciating the intervening state of believing that it's raining, and without appreciating intentions of the speaker that explain the speaker's motivation for uttering this. They apparently have an inability to recognize that someone is lying, and are unable to appreciate that someone could utter something on the basis of a false belief.[89]

When people's utterances disagree with what subjects know to really be the case, Leslie and Roth (1993, p. 96) found that autistic people tend to side with their own knowledge of the way things really are. In one study, autistic subjects were shown a variation of the "Sally-Anne" test, in which Anne lies to Sally about the location of some chocolates which Sally hid, and which Anne subsequently moved while Sally was out of the room. Sally notices that the chocolates are gone, and Anne tells Sally that the dog took them, and they're now in the doghouse. When asked what each person would believe, autistic subjects generally judged that each person believes what the subject themselves knew to really be the case, regardless of what Anne says, or what Sally has been told. They appeared to be unable to deal with the connection between the speaker's beliefs and their utterance, or between what someone is told and their beliefs.

Similarly, Tager-Flusberg (1993, p. 147) notes that early on in their development autistic children make pronoun reversal errors, referring to themselves as "you" and their mothers as "I" or "me", a type of error that non-autistic control-group children never made. Early on, they appear to interpret pronouns to function as names do; apparently interpreting them to be used to refer to a certain person, not as something which functions differently for speakers than for interpreters. A related error in autistic subjects' language-use, was asking questions by employing a questioning word-ordering and rising intonation contour, when "it is clear that they should have been spoken as statements" (Tager-Flusberg 1993, p. 147).

Tager-Flusberg suggests that both confusing pronouns such as 'I' and 'me,' with 'you' and confusing statements with questions is due to the fact that the autistic subjects have not yet figured out the different speaker and listener roles; they produce utterances that should have been spoken by their

[89] In this respect, autistic speakers are living examples of Jonathan Bennett's (1976) "Dullards" (see pp. 194-6). Dullards are similarly unable to recognize a speaker's intention to communicate something, merely recognizing associations between S being uttered and P being true, thus the Dullards cannot "explain unreliability or false utterance as a product of error or insincerity" (p. 195).

conversational partner. This type of error appears to be unique to autistic people. Tager-Flusberg argues that this would be due to autistic people's inability to appreciate that others' have different perspectives on a situation.

Autistic people also appear to often be either unable (or unwilling) to continue a topic of conversation raised by another. Carol Feldman remarks that "autistic speakers seem unable to extend the interlocutor's previous comment. They seem not to know where it is 'going' " (Bruner and Feldman 1993, p. 274). Additionally, when autistic subjects brought up a topic, they had difficulties when their conversational partner replies with topic-maintaining utterances like "Why do you like it?" "What do you do there?" and "What's that all about?" (Bruner and Feldman 1993, p. 277, p. 277). This deficit would be expectable in people who have difficulty grasping the purpose that the other person's utterance was made for. From the particular data that Feldman reports, it seems to be the case that autistic people have particular difficulty with such replies when the interlocutor replies using pronouns; they seem unable to recognize what "it", "there" or "that" is being used to refer to. This could be due to the fact that utterances which employ pronouns are more difficult to interpret literally. Being able to appreciate the speaker's purpose in uttering "Why do you like it?" would be essential to interpreting the pronoun's role in the utterance successfully. For normal language-users, recognizing that the speaker is extending and continuing the topic raised by the last utterance helps specify what the pronoun is being used to refer to. In order to specify the referents of pronouns, the utterance needs to be seen as part of an ongoing interaction, one person's move in the interaction being in response to the other's. A great deal of conversation depends on this understanding of a mutual intertwining of each other's conversational purposes. This purposive element of a conversational interaction seems to be invisible to autistic speakers.

The above types of error could be explained by a deficit in the ability to appreciate a speaker's reason for making the utterance, and an inclination to interpret utterances literally (without an eye to the role the utterance has in the overall interaction). Our linguistic interactions are structured by the different roles of conversational partners, and by the norms governing the turn-about interaction of action, response, and further response, all structured within a shared practice. A speech act counts as a question partly because of the intonation contour of the speaker's utterance, but mostly because of the practice that structures the interaction in which the utterance is set. It's rather expectable that these sorts of errors would be produced, if one of the participants lacks the ability to appreciate this dovetailing of the purposes of each others' linguistic actions and the roles that the norms of the practice allocate to each participant, an to the speech acts (and other actions) they perform.

To summarize this survey, it appears that in autistic people, many of the language-specialized features of the brain are intact. They have some grammatical and lexical abilities, and the abilities to identify heard words and to pronounce words, and so on. However, the mindreading abilities that I'm arguing support all language-use fail to develop. The particularities of autistic people's language-use are explainable as deficits resulting from the inability to recognize others' perspectives and knowledge as different from their own, and especially the inability to encode others' actions as being performed for a reason. These deficits in mindreading abilities, while sparing the functioning of other language-specializations of the human brain, result in a peculiar patter of linguistic deficits. Autistic people who are able to use language have difficulties in recognizing that their own explanations and answers ought to be sensitive to other people's ignorance and perspective. They also have difficulty with interpreting the purpose of others' non-literal utterances, with appreciating the different roles of speaker and interpreter in a linguistic interaction, and with recognizing both the different speech acts people perform and the purposes those speech acts were performed to achieve. It seems reasonable to conclude from this, that these "mindreading" abilities are essential to our normal linguistic interactions. It appears that deficits in these abilities, and these abilities alone, is enough to severely impoverish people's linguistic interactions.[90]

4.6 *Language-use as skilled practice, not application of theory.*

I said earlier that I prefer the term "mindreading" to having a "theory of mind". My principal reason for saying this is that I want to emphasize the extent to which the abilities we are talking about are tacit skills. The contrast between normal speaker's natural sensitivity to other's intentional states, and autistic people's laborious calculation emphasizes this point. The thesis that mindreading is a tacit skill, which underlies our ability to use language, and thus all our practices of attributing contentful intentional states to each other, is another

[90] This is not to imply that there are no deficits in autism other than in the inability to mindread. But it does appear that this deficit accounts for many of the peculiarities of autistic people's social and linguistic interactions.

fundamental tenet of third base approaches to language and intentionality. Yet another way to underscore the third base theorist account I'm proposing here, is the extent to which speaking a language is itself a skill (rather than the application of a theory), based on many other tacit skills besides mindreading abilities.

One way to emphasis this is in terms of a contrast between the "parts" different approaches analyze the performance of a particular speech act into. There are (at least) these two different types of parts: One is to analyze the performance of a speech act in terms of a speaker's intentions to mean something, and their intentions to be interpreted as meaning that by virtue of the expressions they utter, and what contribution their (intended) make to the intended "first meaning" of the sentence uttered. Alternatively, we can analyze the performance of a particular speech act as I have been, in terms of the speaker's intentions to be recognized as performing a certain speech act while (or as a means of getting) engaged in a social normative practice, and by virtue of this, to have certain intentional states attributed to them as reasons for performing that speech act.

Most accounts of people's linguistic skills assume that interpreters concentrate on sentences and their meanings and truth conditions, rather than on the speech acts speakers perform, the practices those speech acts are framed within, and the intentions that are the speaker's reasons for so acting. If theorists do focus on the intentions of a speaker, it is, admittedly, on the intention to achieve a certain response. However, they assume that the means by which this response is produced is by the speaker's intention to *mean* something (as Grice expresses it) or an intention to be interpreted as meaning something (as Davidson argues). My point here is that such accounts assume that speakers concentrate on parts of an utterance –expressions and their meanings– that they do not ordinarily concentrate upon.

Michael Polanyi's (1958, pp. 55-7) distinction between focal and subsidiary awareness, which I explained in the previous chapter, is useful again here. Polanyi's distinction can help illustrate the way people can focus on people's speech acts and reasons, and less on their expressions and meanings. Polanyi argues that in the exercise of skills, the agent concentrates *focally* upon certain aspects of the task, while other aspects are attended to only *subsidiarily*. For example, a blind person using a cane to feel their way concentrates focally upon the objects detected and the path explored, while attending only subsidiarily to the movements of the hand muscles that direct the cane, and the feel of the cane in the hand as the tip of the cane touches things. The elements in subsidiary awareness are subsumed into what Heidegger calls "the background": non-conscious non-intentional capacities and causal mechanisms that are the preconditions of conscious purposeful exercise of skills.

To attempt to analyze or explain the exercise of a skill in terms of particulars that the agent is only subsidiarily aware of, is to perform what Polanyi calls a "destructive" analysis of the skill in question (pp. 50-52, 63). A good example is an expert pianist's "touch" (pp. 50-1). Being able to strike the piano's keys with a certain quality of sound is a distinctive achievement in pianists' skill learning. Yet to ask a pianist to examine and explain their "touch", is to ask them to concentrate on an aspect of piano-playing that they do not normally focus upon. A pianist is typically engaged in playing a particular piece, not in striking the keys with the right "touch". Subsidiary and focal awareness are mutually exclusive, says Polanyi (p. 56). Trying to concentrate focally on particulars that we are normally only subsidiarily aware of degrades performance significantly. "If a pianist shifts his attention from the piece he is playing to the movements of his fingers while he is playing it, he gets confused and may have to stop" (p. 56).[91] A focal concentration on elements that normally play a subsidiary role in the activity, suggests Polanyi (p. 56), is commonly known as "self-consciousness", of which stagefright is a serious form. Such a concentration on the next word, gesture, or note that one has to perform, says Polanyi, "destroys the sense of the context which alone can evoke the proper sequence of words, notes, gestures" (p. 56). This "sense of the context", it would seem, includes the overall practice the agent is engaged in, such as playing the piano sonata, walking to the store without bumping into unseen obstacles, or asking you to close the door.

In a similar way, I maintain that in most everyday, unproblematic cases of the performance of a speech act, our focal attention is on performing a certain speech act (in a way in which the audience can recognize our intention to perform this speech act and the practice it is framed within); we have only a subsidiary awareness of the particular expressions we use to perform the speech act (p. 57).[92] Similarly, interpreters concentrate focally on the speech act the

[91] Polanyi refers here to Henri Wallon, *De l'Acte a la Pensé*, Paris 1942, p. 223.

[92] This is, to some extent, an empirical question, in spite of the theoretical support I see for it. It would be interesting to see (or run) a study to determine the extent to which this is

speaker performs, and only subsidiarily on the expressions the speakers uses. Interpreters may attend only subsidiarily to the actual practice that the speech act is framed within. The practice frames their interaction, but is not consciously attended to. Polanyi gives the example (p. 57) of reading his morning correspondence, which arrives in many different languages. When a letter would be of interest to his monolingual son, however, he says that before showing the letter to his son, he has to stop and check which language the letter is in. While reading the letter, the language the letter was written in was only subsidiarily attended to. His focal awareness was on the content of the letter, not on the particular expressions or the particular language used to convey that content. And when speaking or writing, the skill of selecting the best expression to achieve that objective is similar to a pianist's touch; it is a skill we exercise, but it is one we exercise without being aware of exactly how we do so, and it's an aspect of linguistic actions of which we typically have only a subsidiary awareness.

In cases of breakdown, where what speech act the speaker intends their utterance to count as the performance of is not immediately recognizable, then we may attend focally to aspect that we normally attend to only in a subsidiary way. For instance, in Searle's example of the American soldier the Italians observe that he is speaking German, in an authoritative tone. They cannot recognize the speech act the soldier is performing, but the soldier hopes that they will use the context, his manner of speaking, and the fact that he is speaking German to figure out the practice that he is trying to engage them in. Conversely, when I shout to you "put the wood in the hole" as you leave the door open, you might be unable to concentrate focally on the speech act I am performing, but by concentrating focally on bits of the overall speech situation that you normally would not concentrate focally on –the context, the history of interactions between us, and the particular expressions that I use– you might be able to figure out that I am asking you to close the door. But the fact that we can

true. For instance, it would be interesting to see a discourse analysis of people reporting on conversations that they have recently had, looking at whether people principally report locutionary acts (the particular expressions uttered) or illocutionary acts (the speech acts performed). The ease of explanation and lower demands on memory afforded by the normative practices into which such speech acts are moves, however, would probably bias such a study towards the subjects reporting the kind of speech act performed as opposed to reporting the precise locutionary act. Still, it would be interesting to see if some empirical support could be garnered for this distinction.

do such things in cases of "breakdown" does not entail that we concentrate on these aspects while engaged in normal smooth linguistic interactions.

Thus we perform a "destructive analysis" of the linguistic skills that support people's normal conversational abilities, if we focus on the individual words uttered, what they each mean, and how each word's meaning contributes to the meaning of the whole sentence. This is because the whole act of performing a speech act is analyzed as though speakers attend to *isolated* particulars that must then be combined to produce the meaning of the sentence uttered. But these are particulars of which we are only subsidiarily aware. These words are only meaningful when viewed *jointly*, as a whole (p. 63), says Polanyi, and when viewed in the context of the practice in which they are embedded and its objectives (p. 57). The interpreter's focal awareness is on the speech act performed (framed within a practice); the interpreter is only subsidiarily aware of the words used to perform that speech act. In addition, says Polanyi (p. 63), we originally "gained control" over particulars such as the words used in the performance of a speech act "in terms of their contribution to a reasonable result"; a contribution achieved only when the words are part of whole utterances. Because of this, he says, the words "have never been known and still less were willed in themselves." Therefore to analyze a significant whole (including the context and purpose of the whole) in terms of constituent elements like the words uttered, is to analyze parts "deprived of any purpose or meaning" (p. 63). Because of the "disorganizing" effect of switching attention from the whole to the particulars which jointly make up the whole, Polanyi calls skills such as playing a piano, using a cane, and using words in speech or writing "logically unspecifiable", since "the specification of the particulars would logically contradict what is implied in the performance or context in question" (p. 56).

Yet we do somehow select what expressions to utter in the performance of a speech act. How else could we do this, if not by selecting them on the basis of each expression's meaning? I don't believe this question *has* to be answered in terms of what the expressions used mean. Instead, this can be answered in terms of the speaker's reason for performing the speech act, and the speaker's beliefs about how effectively this purpose is achieved by uttering those expressions, spoken with that tone and inflection, by this person to that person, with that social relationship and history of interactions between them, framed within this practice, with those objects nearby. Different expressions would be used when the speaker had a different purpose in mind, or in a different context.

Selecting the appropriate expressions to use to perform a particular speech act involves what Polanyi calls *connoisseurship*: the discriminative abilities of a skilled practitioner of an art, such as the medical diagnostician or the expert wine-taster (p. 54). Such discriminative skills, he says, "are continuous with the more actively muscular skills, like swimming or riding a bicycle" (p. 54). Connoisseurs of a particular practice have the ability, for instance, to make discriminative selections between actions or tools, based on how suitable or good they are for the achievement of a goal.

For a practitioner of the art of speaking a language, this connoisseurship amounts to the ability to select expressions, based not on their meanings, but on how good or suitable they are for using to achieve the speaker's goals. As Polanyi remarks (p. 63), we originally "gained control" over words in terms of their contribution to goals we were trying to achieve. Thus selection of which expressions to use is done in subsidiary awareness just as the blind person only subsidiarily selects various hand-movements to move their cane as they focally concentrate on exploring their path with the cane. (That is, if they even focally concentrate on that; an experienced cane-user, walking on a familiar path may be focally concentrating on any manner of things while walking along, just as a sighted person often does.) When performing a speech act, the speaker's focal awareness is on the goal of the speech act, and on the way this fits into the overall practice and the overall interaction within which it is framed.

Interpreters, like speakers, are connoisseurs who attend focally to comprehensive wholes, to the speech acts others perform, not to the "parts" that these wholes are constituted from. Interpreters attend focally to the overall meaning of the utterance, Polanyi says (or as I would say it, to the overall purpose of the speaker), and only subsidiarily to the expressions that the speaker uses.

So on this view, speakers are both skilful practitioners and connoisseurs, and interpreters are connoisseurs (who are also skillful practitioners) of the art of using language. They each know how to participate skillfully in interactions and practices that involves linguistic moves. This connoisseurship of both speakers and interpreters is involved in a linguistic interaction. Such connoisseurship, says Polanyi, is

> the pouring of ourselves into the subsidiary awareness of particulars, which in the performance of skills are instrumental to a skilful achievement, and which in the

> exercise of connoisseurship function as the elements of the observed comprehensive whole. The skilful performer is seen to be setting standards to himself and judging himself by them; the connoisseur is seen valuing *comprehensive entities* in terms of a standard set by him for their excellence. The elements of such a context, the hammer, the probe, the spoken word, all point beyond themselves and are endowed with meaning in this context... (p. 64, my emphasis).

I have a small problem with Polanyi's comment that the standard by which the connoisseur values comprehensive entities is a standard set by the connoisseur. It is only partially set by the connoisseur. It is mostly set by the *shared* standards of excellence and appropriateness set by the shared norms of the practice within which the connoisseurship is exercised.

Elements like the spoken words, then, are selected and evaluated only subsidiarily, and only together as a complex whole. They only have meaning in the context of the whole utterance they jointly (but partially) constitute. Whatever meaning each word may be said to have is a logical construction *out of* this comprehensive whole. And the words are only analyzable as having these meanings when part of this comprehensive whole. In other utterances they can be shown to have quite different meanings.

If we're trying to understand the skills involved when people successfully interact linguistically with one another, there's no *need* to bring meanings in. (We may want to talk about meanings when talking about the language itself, and the inter-relationships between its parts, but that's another analytic purpose.) When explaining people's linguistic abilities, I believe it's a mistake to reify meanings, and to talk about them as real entities which people *must* interpret when they converse and understand each other. It's a mistake because expressions only acquire meanings when they are used by a speaker within a norm-governed practice to perform a certain speech act. And the meanings of the words used do not combine to make the meaning of the sentence uttered. Rather the analysis —although I do not endorse such an analysis; certainly not for *all* utterances— goes in the reverse direction: the purpose of the act of making that utterance is the fundamental entity. It can be analyzed in terms of the speaker's meaning, which can then be analyzed to figure out the "meanings" the words used have in the context of this comprehensive whole. (Dictionaries are stuffed full of this sort of thing; lists of the different meanings a word can have, when used in different contexts to do different jobs.) But such an analysis does not represent the kind of analysis that interpreters make when engaged in linguistic activities.

My point here, is that when we see speakers as skillful practitioners of the art of speaking a language (rather than as possessors of a theory about expressions and their meanings) and interpreters as connoisseurs of such practices, we see that the principal focus of both speakers and interpreters is on the speaker's intentions, not on the meanings of the speaker's expressions. The expressions employed and their meanings are elements that we can logically analyze and extract out of an act of making an utterance, but interpreters do not typically analyze utterances this way. Speakers attend focally to the goals of their actions; goals instituted within a shared practice, and achieved by the use of particular expressions. But the practice and the expressions are only subsidiarily attended to by the speaker. (The fact that speakers often fail to notice their own slips of the tongue lends support to this thesis.) Interpreters also attend focally to speakers' reasons for performing speech acts, and perhaps to the further ends these acts are means towards; they attend only subsidiarily to the particular expressions used and to the practice that frames the interaction. To analyze the skills involved in performing speech acts in terms of the selection of words based on what they mean, or the skills involved in understanding speech acts as the interpretation of the meanings of words, and to analyze either in terms of a compositional account of how these word-meanings contribute to the meaning of the whole utterance, is to perform a "destructive analysis" of the linguistic skills involved. Such an analysis casts the skill of making and interpreting utterances in terms of particulars that play little role in people's focal awareness as they exercise that skill.

However, an analysis such as the one I gave in the previous section, in terms of the purpose of the utterance, and the speaker's intention to get the interpreter to recognize that purpose, I want to argue, is closer to an analysis in terms of the wholes that the speaker and interpreter attend to focally. Analysis in terms of words and their meanings may play an important role in certain kinds of analysis of certain kinds of sentences and their truth-conditions. However, such an analysis is not usually appropriate for explaining the kinds of skills and abilities people employ in their everyday linguistic interactions. More appropriate here is an analysis in terms of speakers' intentions to perform particular speech acts and to engage in certain norm-governed practices, speakers' intentions to get interpreters to recognize their reasons for so acting, and interpreters' attributions to speakers of intentions to perform certain speech acts.

4.7 Successfully dealing with "slips": a speech act based explanation.

Our ability to successfully deal with slips of the tongue, malapropisms, unfamiliar jargon and the like, then, is an extension of the basic human ability to participate in practices and to interpret people's actions as signs of their intentional states. We almost have no choice in interpreting people's actions as purposeful; it's almost an involuntary reflex to look for the purpose driving people's actions.

We can also recognize what someone tried to do when they don't successfully pull off the act as they intended. A recent study by Andrew Meltzoff (1995) has shown that this ability to tell what someone is trying to achieve, when they don't achieve it, is one that even eighteen month old infants possess. Capitalizing on infants' tendency to pick up behaviour from adults, and to re-enact or imitate what they see, Meltzoff showed eighteen month old infants an adult trying to perform a certain action, such as pulling apart a dumbbell, or hanging an elastic band on a hook, yet failing each time to do it. He then allowed the infant to play with the equipment the adult had been using. The infants, in a significantly higher number of cases than control groups, successfully re-enacted the action the adults had attempted to do, in spite of the fact that the infant hadn't seen this action being performed successfully. From this experiment, and similar ones I won't go into here, Meltzoff concluded that "18 month old children can understand the intended acts of adults even when the adult does not fulfill his intentions".

In the experimental situations within which Meltzoff's subjects demonstrated these abilities, the situations were not social ones, that could be framed by social norms and the expectations those norms engender. However, such situations and norms could only aid interpretation of the adult's intentions. When we also consider the practices that people are socialized into knowing how to participate in, and how their norms structure the kinds of things it is appropriate to do in certain situations, it's easy to see how people can become very sensitive to the kinds of things people intend to do. We become very skilled at positing intentional states as reasons for people's actions, even when they fail to achieve their goals.

The sort of ability enables people to cope with linguistic "slips" like those Davidson brings to our attention. People are able to successfully cope with "defective" utterances like when my partner uttered, "I'm late! I have to go *now*, but I can't find my key-cars," or when my grandmother said that there's so many "turners to corn," or to deal with unconventional utterances like my roommate asking "Can you grab the wheel-thingummy?" To do this, the audience needs to recognize, not as Davidson supposes, what the person making the utterance *intended their expressions to mean*, but what the person making the utterance was *trying to do* in uttering that. The speaker intends the audience to recognize their intention to engage in a certain practice, by performing a speech act that counts as a move in that practice. And the speaker intends that, on the basis of this recognition, the audience recognise the further intention to get the interpreter to respond appropriately (as the norms of that practice dictate) to the performance of that speech act. My partner uttered this in order to get me to help her find her car-keys; my grandmother was criticizing a set of directions more complex than they needed to be as a preface to offering an alternative; my roommate was asking me to hand her the pizza-slicing wheel so that she could use it to cut the pizza. In each case the speech act is performed in order to achieve some interactive purpose, by engaging with the interpreter in an interaction framed by a shared set of norms for that kind of interaction.

The person spoken to can understand the utterance because they can still recognize what speech act the speaker was trying to perform in making this utterance. Sometimes the speech act the speaker intends to perform isn't immediately recognizable, but in such cases the interpreter can attend focally to aspects of the speech situation that are normally only subsidiarily attended to. Thus they can perhaps recognize the practice that the speaker is engaged in or the practice that they are trying to engage in. Thus, by virtue of the norms of this practice and the kinds of ends one can achieve by engaging in that practice, the further end to which this speech act was a means can become recognizable. This allows after-the-fact reconstruction of what speech act the speaker intended their utterance to constitute the performance of. It's this that facilitates "correct" interpretation; that is, it facilitates the interpreter's ability to respond appropriately.[93]

[93] Indeed, I often find myself not being able to figure out exactly what speech act a conversational partner performed, but because we're, say, in a noisy bar where this happens a lot, one cannot ask a conversational partner to repeat everything they say. In

In cases of "breakdown" like this, interpreters are also likely to attend focally to the particular expressions used, and to use these as additional clues to the kind of speech act the speaker intended to perform. But because they are cases of "breakdown", the words do not sit together as a unified whole. They must be examined piecemeal, in combination with the other clues available, such as the social and physical context of the speech act (what practices we are engaged in or what practices the speaker is likely to be trying to engage me in, what their speech act was a response to, what speech acts are expectable in the present situation, and so on). The purpose of such examination, however, is not to figure out what the expression in this case means, but to figure out what speech act the speaker intended that utterance to count as the performance of. I might then try to figure out what expression the speaker intended to use, or what expression I would have used. This, however, is usually done *after* I first figure out what speech act the speaker intended to perform.

Davidson himself comes close to noticing the role of employing means to ends relationships in figuring out how to interpret a non-standard utterance, when he discusses the sonnet by Shakespeare that includes the line: "On Helen's cheek all art of beauty set/ and you in Grecian tires are painted new". He says that we "can descry the literal meaning of a word, by first appreciating what the speaker was getting at" (p. 435). He notes that the

> "intentions with which an act is performed are usually ordered by the relation of means to ends... Thus the poet wants (let us say) to praise the beauty and generosity of his patron. He does this by using images that say the person addressed takes on every good aspect to be found in nature or in man or woman." (p. 435)

So we can understand the word "tires" in the sonnet, if we recognize the poet's ultimate end: he is trying to invoke the practice of praising the beauty of his patron. Within this practice, he uses a certain locution to achieve his aim of comparing the patron to Helen of Troy, effectively performing the speech act of praising the patron's beauty by saying that seeing Helen in all her beauty is like

such cases, however, I am often able to figure out enough of the context of the speech act and the practices we are engaged in to determine a somewhat appropriate response anyway ("Uh-huh" often suffices). I respond in this kind of way with the hope that it will be appropriate enough that it will not cause the entire practice to break down, and that not interpreting that speech act will not have serious consequences.

seeing the patron in Grecian attire. In order to do this, says Davidson, the poet uses "tires" as though it has the role in our language that "attire" normally has. To Davidson, Shakespeare constructs a situation in which the interpreter must construct a passing theory of meaning in which "tires" has the role that "attire" has in our prior theory.

But after noting that we can understand this unfamiliar phrase by "first appreciating what the speaker was getting at," (Praising the patron's beauty), Davidson diverts from this insight, stating that he is interested in *the intention to be interpreted in a certain way*, which he sees as the *first* means to the speaker's aim; this intention to be interpreted in a certain way is the intention that specifies "first meaning".

This is supposed to be the case even for normal interactions where there is no "breakdown" nor any unusual uses of expressions. Let's return to Davidson's example of Diogenes uttering to Alexander the Great (the Greek equivalent of), "I would have you stand from between me and the Sun." To Davidson (p. 435), Diogenes intends three things, each as a means to a further end. First, Diogenes intends that Alexander interpret him as *meaning* that he would have Alexander move from between him and the Sun. Alexander's recognizing this first intention is intended to be a means of achieving a further end: that of getting Alexander to recognize that Diogenes is *asking* him to move. Diogenes furthermore intends that Alexander's recognizing this further end, function as a means of *getting* Alexander to move. This first intention, the intention to be interpreted as *meaning* that he wants Alexander to move, specifies Diogenes' "first meaning", and is what Davidson sees as the first step in the series of means to ends that Diogenes employs.

I see this as mistaken, principally because this first intention Davidson notes, the *intention to mean a certain thing,* is an understandably alluring abstraction, but (at least for present analytical purposes, in trying to understand the mechanisms underlying people's linguistic and interpretive abilities,) an ultimately vacuous abstraction. The analysis of people's skills in making and interpreting utterances in terms of sentences and their meanings is (as I argued in the previous section) an analysis of the skill in terms of particulars that play little role in the focal awareness of the people exercising the skills analyzed. What the speaker *intends to mean* is not something that people focally attend to when they converse. If they do attend to this, it is only in a subsidiary way, and in cases of "breakdown" when an after-the-fact reconstruction of what happened is required.[94]

So Davidson has the process inverted. Davidson sees the intention to be interpreted as meaning a certain thing as a speaker's first intention. One of Davidson's principal difficulties, however, is (or should be) explaining *how* we are able to interpret someone's expressions as meaning what they intended them to mean, when their expressions don't normally mean that. We should find it rather difficult to recognize someone's intention to make their expressions *mean* something that they don't standardly or obviously mean. This would require a more telepathic form of "mindreading" than the kind I've been talking about. Davidson's explanation of *how* we are able to interpret someone as meaning what they intended to mean, when their expressions don't standardly or obviously mean that, is rather unilluminating (see Davidson 1986, p. 446).

It's much easier, however, to explain how we are able to understand slips of the tongue and other non-standard utterances if we approach this in terms of our ability to recognize a speaker's intention to *do* something; especially when their action was set within a recognizable practice. In fact, first recognizing what the speaker intended to do and/or the practice that their speech act is intended to count as a move within, can often provide the clue which enables Davidson's interpreter to figure out what meanings the speaker intended their expressions to have. For instance, if I point towards the open door that you have just left open, and utter "Put the wood in the hole", what I intend to be interpreted as *meaning* isn't all that important here. What matters to me, is that you recognize my reason for doing that: I'm trying to engage you in the practice of asking you to close the door. And since the phrase I'm using to do this is a rather unconventional tool, which doesn't make my intention to ask you to close the door as recognizable as it could be, it might only be by first recognizing that I intended to *ask you to close the door* (by also noticing the open door and my gruff tone of voice), that you could subsequently come to a "correct" interpretation of the meaning of the expressions I used and to an appropriate response to my speech act. You could do this by subsequently considering focally the particular

[94] As often happens in philosophy, cases of breakdown —where special means are operative— are taken to be indicative of the means used in ordinary cases where nothing goes wrong. For example, a theory of sense data is the result of taking the exceptional situations of dreaming and hallucinations to indicate the mechanisms underlying perception in general. (See Chapter Two for a criticism of the representational theory of perception that results.)

expressions I used, which you initially attended to only in a subsidiary way. (I say "subsequently" because, as Polanyi (1958, p. 56) rightly argues, we cannot attend to the same aspect of an action in both a focal and subsidiary way.[95]) Because of this ability to perform an after-the-fact reconstruction, the interpreter can focally consider the particular expressions I used, and after recognizing my purpose and using it as a clue, they can (to put it in Davidson's terms) construct a passing theory in which "Put the wood in the hole" is given all the powers, relations and roles that "Close the door" has in their prior theory for interpreting my utterances.

It is easy to see how people can recognize the reason that someone performs a certain physical action, even if they don't pull off the action as they intended. Clues come from the physical, historical, psychological and social context: from whatever practices people conventionally engage in, in such contexts, from expectable goals people aim to achieve in such contexts, and from the means conventionally used to achieve those ends. It's similarly easy to recognize what someone tried to do linguistically; especially since it is usually the speaker's intention to make it easy for you to recognize what practice their speech act was supposed to be a move in, and what kind of move it was supposed to count as making. Even if someone makes a slip and doesn't perform the linguistic action as they intended, or if they don't make their reasons for making this utterance as immediately recognizable as they intended, we can still figure out likely candidates for their reason for doing as they did. Thus we can use these likely candidates for what the speaker was trying to do, to engage the speaker in that activity, and to show the speaker that they have secured "uptake" by producing a response we think is appropriate.

4.8 "There is no such thing as a language."

A further point in favour of my "third base" approach to language and to the capacities that support our language-use, is the pessimistic conclusions Davidson

[95] Hofstadter's elegant and illustrative "Prelude, Ant Fugue" (in Hofstadter (1979) and reprinted in Hofstadter and Dennett (1981)) makes a similar point, that how one cannot simultaneously attend to a whole and to the parts the whole is made up of. While talking (in the voices of Achilles and the Anteater, p. 156-8) about listening to a fugue, he makes the point that while attending to one voice in the fugue, the others, and the whole that the voices comprise is also present, but not attended to. While listening to the whole made up of the combined voices, we attend subsidiarily to each voice, but have to listen to it again to attend to how the one voice contributed to that whole.

himself draws about "theory of meaning" approaches to language. To Davidson, when interpretation is successful (especially in cases where expressions are used in unusual ways), interpreters share with speakers the ad hoc passing theories of interpretation they construct at the time of interpretation (to interpret this utterance only, of this person only), not their pre-learned prior theories of interpretation. So when interpretation is successful, the theory of the meanings of words that language users *share* is not *learned beforehand,* and the theory of the meanings of words they have *learned beforehand* is not *shared.* This disagrees explicitly with the three principles I cited at the beginning of this chapter, about the theory of the meaning of people's utterances that standard theories of language take as fundamental. To what Davidson calls "standard theories" of language, meanings are *systematic, shared,* and *pre-learned.* Davidson concludes from this that there is no such thing as a language, if what we mean by a language is the kind of thing philosophers of language and linguists have assumed it to be. A language can't be "a clearly defined *shared* structure which language-users *acquire* and *then* apply to cases" (p. 446, my emphasis).

I agree with Davidson's conclusion that what people learn when they learn a language can't be the sort of structure that many philosophers and linguists have supposed it to be; not if this structure is a theory of the meanings of words and the systematic way that word-meanings combine to create the meanings of sentences uttered. Human linguistic abilities are not well accounted for by theories of language which explain these abilities in terms of a theory of interpretation by which language users interpret the of word-meanings that are shared and known in advance, and construct sentence-meanings from them.

Davidson argues for an alternative account of what speakers and interpreters share: they must share the ability to construct successful (that is, convergent) passing theories of interpretation. Interpreters construct such theories on the fly, in a very ad hoc fashion, out of their prior theories. A passing theory of interpretation, is for use on this occasion only, to interpret this utterance only, of that speaker only. The fact that, to Davidson, to interpret someone's utterance we must construct an *ad hoc* passing theory of meaning for this person only, on this occasion only, to interpret this utterance only, counts heavily against his approach. To Lakatos (1974), that *ad hoc* theories or modifications to theories like this are required to account for phenomena, is a sign of a degenerating research programme.

I'd take Davidson's conclusion even further, then. Language cannot be anything like what the standard theories suppose, and making ad hoc modifications to the standard theories seems to be a last attempt to prop up a set of theories that are fundamentally flawed. If an ad hoc explanation like this is the best way to explain people's linguistic abilities from within a theory of meaning approach, then it seems that we should stop attempting to explain people's linguistic abilities in terms of the production and interpretation of meanings. The alternative approach to explaining how it is that people are able to use language that I'm advocating, by explaining this ability in terms of people's general ability to recognize the reasons for people's actions (by appealing to the practices that frame the actions, and the norms of such practices), accounts for this ability in a much less ad hoc way than Davidson's apparatus.

Additionally, I don't agree with Davidson's conclusion that the principle that language use is "governed by learned conventions and regularities" cannot stand (p. 446), nor with his pessimistic last recommendation that

> We should try again to say how convention in any important sense is involved in language; or, as I think, we should give up the attempt to illuminate how we communicate with one another by appeal to conventions (p. 446).

Davidson ignores the purposefulness and the practice-situated nature of people's linguistic and non-linguistic actions, and instead looks for conventions at the abstract level of the meaningfulness of the words uttered. Thus, he expects that if we are going to find rules and conventions anywhere, we should find rules and conventions *governing the meanings of words*. Because Davidson doesn't find shared, pre-learned rules and conventions there, he concludes that we won't find them anywhere. Therefore, he says that we should abandon the notion that rules and conventions underlie language-use, and we should abandon the notion of language as something that is both pre-learned and shared.

I'd prefer to take the former of the two alternatives Davidson presents in the quotation above, and abandoning an account based in conventions governing the meanings of expressions, try to say again how convention (normativity) is involved in language-use. After Davidson concludes that the principle that first meanings are governed by learned conventions and regularities cannot stand, he goes on to say that "it is unclear what can take its place" (p. 446). It is perfectly clear what can take its place, however, if we stop viewing language in terms of interpretation of "first meanings" according to a (prior or passing) theory of interpretation. Instead, we should view language-use as people skillfully performing purposeful actions as moves within social practices, while "advertising" the purpose that is the reason for their action. If we do so, we can find the shared and pre-learned rules and conventions that underlie our linguistic abilities, not in shared and pre-learned conventional meanings of expressions, but in the shared and pre-learned norms of the practices that constitute our shared form of life.

To learn a language is to learn how to participate skillfully in these social practices; a skill all users of the language share (although no-one is familiar with *all* such practices). For example, most social practices, including linguistic practices, are bound by conventions of what sort of acts are expected and what acts would be inappropriate, following particular acts by others. Winograd and Flores' diagram, which I gave in section 4.3, of the legitimate linguistic moves in the practice of A asking B to do something, is a good illustration. Note also that much of this know-how is rather tacit and unarticulable by ordinary language-users; for instance, linguists expend much conceptual and analytical effort to prise many of the grammatical norms out of languages and make them explicit. The legitimate moves that Winograd and Flores' diagram lays out are also moves we are able to recognize as legitimate, and are able to see as open options when we are participating in this practice, but most of us would have had to expend considerable analytical effort in making them explicit as Winograd and Flores have. Because of such explicit and tacit norms of what kinds of responses are opened by one person's speech act, when that act counts as a move within a certain practice, we have certain expectations of what type of actions appropriately follow. These expectations assist greatly in recognizing the intentions of participants' subsequent linguistic actions. Many of the general shared and pre-learned theories that people employ in interpreting one another's utterances, are theories of this sort: theories about what actions are appropriate following a conversational partner's actions, rather than theories about the meanings of particular expressions.

The effectiveness of our speech acts also depends upon conventions governing the use of particular expressions. For a speaker's use of words to be an effective move within a practice –to count as making that move in that practice– people must recognize that the expressions of an utterance are being used for that purpose. So uttering recognizable expressions in way that the norms of the practice deem to be appropriate is the surest way to achieve your purposes. As Donnellan (1969) points out, you cannot intend to achieve something by a certain means, unless you believe or expect that the means you use will, or at least could, achieve the desired outcome. I can only use particular expressions to achieve a purpose –that is, to perform a certain speech act as a move within a practice– that I believe or expect my audience can and will recognize I intend to achieve. Particular expressions have conventional uses, and we rely on these conventions of use to make our purposes recognizable to others, and hence to make our speech acts effective. Words are tools we use to achieve our purposes in performing speech acts. And they depend on conventions of use for their effectiveness.

This harks back to Heidegger's distinction between ways an object can be used and ways that it should be used. The expressions of a language are special cases of tools. Each expression is *for* performing certain norm-governed actions (expression are special case, it depends on being *recognized* as being used for that purpose in order to achieve that purpose). Within the context of human practices and the norms that permeate our shared forms of life, particular expressions count as being for particular tasks. It's a *connoisseurship* (in Polanyi's (1958, p. 54) sense) of the skilled practitioner of language that enables a speaker to select the appropriate tool for the task at hand. Physical tools *can* be used for all kinds of tasks that the norms of the "equipmental nexus" specify as not what they *should* be used for (I can use a screwdriver to carve my name in the picnic table, or use a crescent-wrench to hammer in a nail). This distinction also applies to expressions. Using unconventional expressions *can* still get the job done, but this is probably less efficient or reliable. Shouting "put the wood in the hole" is probably less efficient and reliable in contexts where this phrase is somewhat unconventional, because it's less conventionally recognizable as a request to close the door, than saying "please close the door." And familiar expressions *can* be used in unconventional ways and still get the job done. For example, a supervisor once recommended to an electrician colleague of mine: "the cabinet's so close to the wall on the right that you'll have to go in from the left and strip the wires behind the cabinet with your left hand. He replied "I think I'm amphibious enough to do that". The word "amphibious" *should* not be used in this way, but in spite of this, its (mis)use in this way *did* still get the job done. The interpreter of this expression can still recognize the speech act in which it was used as the move of accepting the recommendation (probably because of the practice within which it was interpreted as a move). It's the linguistic equivalent of using a chisel to tighten a screw. The tool can be "blunted" and become unsuitable for its normal job if it's misused too often. It may eventually even become only good for the parasitic use to which it was co-opted. (The etymology of many words seems to display this pattern.)

So on occasions when someone makes a slip of the tongue or uses unfamiliar jargon, we can understand the reason they did what they did, and can recognise the appropriate responses open to us, because of these social norms governing the practices we conventionally expect people to engage in, in certain contexts, norms within those practices governing the legitimate moves open to participants, and norms governing the uses expressions should and should not be put to in order to perform speech acts as moves within such practices.

In summary, then, the speech act approach I have outlined here explains our ability to successfully deal with the kinds of linguistic "slips" Davidson brings to our attention, in a much smoother way than the *ad hoc* explanations in Davidson's meaning-based approach. Here we have an approach to language in which we interpret not the meanings of linguistic expressions, but people's actions: speech acts and other actions that are performed with the intention of making a particular move within a norm-governed practice. Our ability to understand people's utterances rests, not on the ability to interpret utterances' meanings, but on the ability to attribute intentions to others when we observe them acting that serve as their reasons for performing that action. These skills of recognizing and attributing purposes to people, are based on our ability to participate in, and familiarity with participating in, interactive norm-governed social practices. Attributing purposes to others requires familiarity with the practices certain contexts evoke and the moves people can legitimately make within certain practices. It also requires a connoisseurship of the skillful practice of appreciating the uses expressions can be put to and that of selecting expressions that are good or suitable for a particular linguistic purpose. All of these abilities are easily explicable and learnable; we learn them by immersion and apprenticeship, by participating in the social practices we end up using them in, under the guidance and following the examples of those already proficient in these practices and skills.

In the next chapter, I will discuss another sort of norm governing our linguistic practices and the moves we make within them. Austin's *felicity conditions* are norms that an action must meet in order to count as a speech act of a particular type. These norms govern the *felicity,* or happiness of certain actions performed within social practices. A warning counts as a fair warning, an estimate counts as good or a true one, a promise counts as a sincere promise, all according to these norms. I will argue in the next chapter that these felicity conditions also govern the criteria that must be met for an attribution of an intentional state to be felicitous. These criteria, I will argue do not involve internal states of the person the intentional state is attributed to, but publicly observable aspect of the situation in which the intentional state is attributed: particularly the actions of the person to whom the intentional state is attributed.

Chapter Five

The Normative Practice of Attributing Intentionality

> *"When I use a word" Humpty Dumpty said, in a rather scornful tone, "it means just what I choose it to mean—neither more or less."*
>
> *"The question is," said Alice, "whether you can make words mean so many different things."*
>
> *"The question is," said Humpty Dumpty, "which is to be master—that's all."*
>
> —Lewis Carroll (1871)

5.1 *Austin's Felicity Conditions.*

Using symbolic languages to perform speech acts is a paradigm example of the kinds of norm-governed interactions human beings are capable of. A speaker should use particular expressions in ways that conform with shared norms of how one should use them, if they want their interpreter to be reliably able to recognize what speech act they intend to perform. When speaking with one another, people assume that that their conversational partners are aware of, and working within, the same normative structure as themselves. This normative structure that all members of our forms of life are socialized into, undergirds every linguistic interaction. It supports the speaker's expectation that using these words, in this situation, will be recognizable to the hearer as the performance of that speech act. It also supports the hearer's ability to interpret the speaker as performing that speech act.

The felicity conditions on the performance of a speech act (Austin 1962) are another important part of the kinds of interdependent normative relationships characteristic of third-base accounts.[96] The norms I talked about at

[96] These felicity conditions, being a set of interlocking norms that undergirds all human language use, is one reason that I'm reluctant to characterize Austin as a second base theorist (that's where Haugeland (1990) puts him; p. 422, note 16). He appears to be a case, like Davidson and Dennett (see Haugeland 1990, p. 418) of someone who

the end of the last chapter govern how one ought to use words, and what moves are appropriate at a particular stage within a practice. Felicity conditions govern whether a purported move or an attempted move genuinely does count as the performance of that speech act, as that move in the practice, and whether that move was an appropriate to make. (Some people refer to a special type of such appropriate moves as "true" speech acts.) Some of these felicity conditions are constitutive of being a certain type of speech act. An action counts as the performance of a speech act of that type, if these conditions are satisfied (by the performance, by the performer's identity and their subsequent and previous behaviour, by the interpreter's behaviour, and by the state of the world). Other felicity conditions must be met for that speech act to count as successful, or valid, or sincere and not "hollow" or "void".

Austin's felicity conditions could be used to rule on some of the other normative aspects of linguistic interactions that I talked about last chapter. For example, the norms of using referential words as they should be used (using them to refer to the things that they should be used to refer to) are also part of the conditions a speech act should meet. The relations between speech acts and other actions the speaker or audience perform are another important part of the set of interdependent normative relationships that govern speech acts. They relate to conditions of securing uptake, and the condition that the performance must be executed completely; "completely" may involve further behaviour on the part of one of the participants in the interaction. For example, orders are related in this way to actions of obeying the order; promises are related to the actions that fulfill them; apologies are related to the actions apologized for, and to acceptances of apologies; bets are related to acceptance of bets and to paying up on lost bets; requesting someone to do something is related to their actions of offering to do it soon, suggestions of alternative actions, performance of the action requested, refusals to perform the action requested, and to reports of the action's having been performed. These relations with other actions could perhaps be covered by some of the felicity conditions (e.g. the conditions that the procedure being executed completely, and that the participants actually conducting themselves in the future according to the procedure), but even if so, these relations are worth pointing out explicitly. In spite of these ways felicity

incorporates elements of two different bases. But unlike Dennett and Davidson, who straddle first and second bases (although Davidson also incorporates elements of third base), Austin would seem to straddle second and third bases.

conditions could cover the kinds of norms I have already been talking about, there are some extra dimensions of normative assessment of speech acts that they bring out, that are especially relevant to my eventual goal of looking at a particular kind of speech act. I'm leading up to talking about the conditions that make the speech act of attributing intentional states to others felicitous.

One noticeable difference between these normative relations and the "equipmental" normative relations Heidegger talks about, is that these norms govern not *items,* as they do on most interpretations of Heidegger,[97] but *actions* (including but not limited to actions of performing speech acts using linguistic items). To Heidegger this object *is* a bow, and is related to arrows and to targets, spears, crossbows, archery halls and quivers, because of the norms governing bows and their "equipmental" relations. To Austin, however, certain words, such as "I hereby promise to..." are related via norms of word use to actions of promising, but the words *themselves* do not constitute a promise. They are for making promises (and derivatively for reporting on promises made, etc.). But, importantly, it's the *action* of uttering words like these that is governed by felicity conditions. This *speech act* is a promise, and is related to actions of carrying out the action promised (among other things), by virtue of the felicity conditions on the practice of promising.

This slight change in focus, from norms governing relations between *items,* to norms governing relations between *actions,* is going to be very important later on. I'm going to argue that it's being set within the context of norms like these —especially norms governing speech acts— that makes any human action, including non-linguistic actions and "mental" actions, the type of action that it is. Importantly, these same norms also relate attributions of people's intentional states ("Mason believes that *p*" and "Mason is thinking about *q*" and so on) to the states of affairs that make such reports true —or rather those that make them felicitous.[98] These states of affairs, I will argue, are also actions; the actions of the person to whom the mental states are attributed. The person's actions license or justify the attribution, and their actions also rule the attribution infelicitous.

[97] Haugeland (1990) p. 408-9; Dreyfus (1991), p 151: "Equipment displays generality and obeys norms."

[98] As we all see soon, the dimension of "fittingness to the facts" (a dimension along which the assessments of "true" and "false" are at opposite ends) is one dimension among many of an utterance's felicity. See Austin (1962), pp. 139-48.

Because the felicity conditions on speech acts are going to be centrally important to my overall argument for this point, I need to spend some time now outlining them in detail. It will be important to have a good picture of the kinds of conditions that our speech acts are assessed according to, when I come to talk later on about the felicity (appropriateness) of asserting "Mason believes that p". The following account of felicity conditions I'm proposing is a possibly contentious interpretation of Austin's position. I do not wish to get into textual interpretation and defense here. I have drawn this account from Austin's account in *How To Do Things With Words*. However, it is defendable in its own right as a sensible view to hold, even if some might disagree about whether Austin held precisely this view.

Looking at the many ways a speech act can fail to be felicitous illuminates the conditions by which we judge speech acts, the conditions that felicitous speech acts accord with. For Austin (1962, chapters 2-3), speech acts can fail to be felicitous for many reasons. Austin divides the many ways speech acts can go wrong into two broad categories: *misfires* and *abuses* (p. 14-18). Misfires occur when we say that the speech act failed to be successfully pulled off for some reason. Abuses occur when we say that the act is professed to be pulled off successfully, but is hollow or insincere; an abuse of the procedure. Misfires are further subdivided, into (A) *misinvocations* and (B) *misexecutions* (a.k.a. miscarriages). These two, combined with (C) *abuses*, comprise the three categories of infelicity, each of which contains several felicity conditions. In what follows I present Austin's names of types of infelicity, and the formulation of the conditions, as Austin presents them, that a speech act must meet to avoid that type of infelicity. I also list some examples to flesh the infelicities out. The conditions are not intended to be completely distinct, nor exhaustive. Many cases could be ruled infelicitous by appeal to either of two conditions (e.g. p. 35), and there may well be types of infelicity not covered here.

Part A: *Misinvocations;* Act unsuccessful

(A.1.1) Non-plays: *There must be an accepted customary procedure for performing the act, the procedure to include the uttering of certain words by certain persons in certain circumstances.*[99]

[99] I have separated A.2 and C.1 into parts, to illustrate the different types of infelicity that Austin groups under this name. A.1.2 was not originally in Austin's list of types of infelicity, but he refers to failure of uptake often as a special type of infelicity. It seems best to fit at A.2.1

In order for me to make a bet with you, I depend on you and I accepting the practice of using words such as "I bet you that..." to make bets. Conversely, there is no longer an accepted procedure whereby in the performance of a speech act I can challenge you to a duel to the death. You could felicitously shrug off such a purported challenge, since our community no longer accept the procedure of challenging people to duels (pp. 26-34).

(A.1.2) *Misunderstandings, failure of uptake: The participants must understand which procedure is being invoked.*

I do not felicitously place a bet with you if you take me to simply be making a prediction. I do not felicitously assert something if you take me to be guessing (p. 22, 138).

(A.2.1) *Misapplications: The circumstances for performing the act must be appropriate.*

The captain of a ship cannot marry a couple ("I hereby pronounce you husband and wife") unless the ship is at sea. I cannot marry a woman ("I do") unless we have a license and are in front of a marriage celebrant. The wedding must be a genuine one, not on a stage in a play. I cannot give you something ("Here, have this; it's yours now") if it's not mine to give (p. 34). I cannot place a bet on a certain horse in race three, if that horse isn't running in race three. Warning you that the bull is going to charge is infelicitous if the bull is not going to charge (p. 55).

(A.2.2) *Misplays: The persons performing the act must be appropriate.*

I do not place a bet if I am six years old (or I can be excused for not paying up if I lose, for this reason). The purser of the ship cannot marry a couple. I do not name a ship by smashing a bottle of champagne over the prow and kicking the chocks away ("I hereby name this ship the *Generalissimo Stalin*"), if I am not the person appointed to name the ship. I cannot order you to do something if I am not in the appropriate position of authority (p. 137). If you inform me that you are in a good mood today, I cannot felicitously state that you are wrong about that (though I might conjecture or guess or argue this) (p. 137).

Part B: *Miscarriages;* Act unsuccessful

(B. 1) *Flaws: The procedure for performing the act must be executed correctly*

Uttering a malapropism, saying something I did not intend to say (e.g. Slips of the tongue; asserting "the bat is on the mat" when I intended to say "cat") (p. 137-8). Employing vague formulas, such as referring to "my house" when I have two of them (p. 36). Similarly for using expressions that disallow other terms: "The house was painted green; we used different shades of red paint for the walls, trim and roof."

(B. 2) <u>Hitches</u>: *The procedure for performing the act must be executed completely*
I don't felicitously make a bet with you if you do not accept my bet (the bet is "abortive"; p. 36). I do not you marry unless both of us say "I do" at the appropriate times and the marriage celebrant pronounces us husband and wife. My attempt to ceremonially open a library is abortive if I say "I hereby open this library" but the key breaks in the lock (p. 37). There are also questionably abortive cases, such as when I give you something by telling you, "Here, have this. It's now yours." but fail to hand it over (p. 37).

<u>Part C: *Abuses;* Act successful but "hollow"</u>

(C.1.1) <u>Insincerities</u>: *The participants must have the requisite thoughts, beliefs and feelings*
My stating something is infelicitous if I do not believe it. ("The cat is on the mat but I do not believe it.") My congratulating you is insincere if I do not at all feel pleased for you (p. 40). My forgiving you is insincere if I do not feel appropriately towards you (a lack of resentment, perhaps?). My promising to do something is infelicitous if I do not believe that it is possible for me to do as I promise.

(C.1.2) <u>Mistakes</u>: *These thoughts, beliefs and feelings must in some sense fit the facts*
My stating something believing it to be true, is infelicitous if this belief is incorrect. My giving you something that I believe is mine, is infelicitous if it is not mine to give (p. 42). My promising to do something that I believed that I was able to do, is infelicitous in the same way if I am not in fact able to do what I promised. (Austin remarks that these infelicities are a special kind; they do not make the act void, but make it excusable; p. 42). My forgiving you is a mistake if you have done nothing for which you need to be forgiven.

(C.1.3) <u>Insincerities (again)</u>: *the participants must have the intention to conduct themselves in the future according to the procedure*
My promising is infelicitous if I do not intend to do as I have promised. Placing a bet with you is infelicitous for this reason if I do not intend to pay up if I lose. My welcoming you is insincere if I intend to treat you as

an enemy (p. 44). My forgiving you is insincere if I intend to keep reproaching you for your misdeed.

(C.2) *Breaches: the participants must actually conduct themselves in the future according to the procedure.*

My assertion is infelicitous (breached) if I do not back it up (provide evidence) if challenged. My promise is similarly breached if I do not do as I promised to do. Debatably, my entreating you to do something rules "out of order" my protesting at your doing as I entreated (p. 44). My treating you as an enemy "breaches" my welcoming you. My asserting something is breached if, when faced with questions or objections, I fail to offer any defense of or evidence for the truth of what I asserted.

Conditions C.1.1 and C.1.3 are important for my purposes. This type of condition on the felicity of a speech act –that the participants in a speech act have the requisite feelings, beliefs, thoughts and intentions– opens up the door for questions about what makes it the case that the participants do indeed have such intentional states. I do not disagree that this is an important condition on the felicity of a speech act. However, there are important concerns about what "having the requisite thoughts, feelings, beliefs, and intentions" amounts to.

5.2 *Brain states cannot be truth-makers for attributions of intentional states.*

These considerations are closely related to the above set of felicity conditions. Like Austin, first base theorists such as Fodor, Cummins and Dretske, and Searle outfield of them, and second base theorists like Bennett, Quine, and Dennett, and folks who straddle the two bases like Davidson, would all take criteria like C.1.1 and C.1.3 to be met by the agent in question actually having the requisite intentional state. There are disagreements about what would have to be the case for this to be so, however. Many of these theorists (especially the first-base theorists) resort to a form of essentialism when talking about humans and their mental states. They believe that there must be a fact that would discern whether or not the system *really* has the intentional states, over and above the behaviour that justifies the attribution of the intentional state (and also counterfactually, the behavior the system would produce, given other circumstances). This fact is usually taken to be a fact about the individual to whom the intentional states are attributed; for example, a neurological fact or a fact about their "mind".

Searle is a good example of this view. He is convinced that an ascription of intentionality must be either true or false, and what makes it so (what we could call the *truth-maker* of the ascription) is something over and above the way the system behaves. Searle claims that something could behave in ways that license attributions of intentionality to it (ways that are only explainable from the intentional stance), but it could still lack genuine intentionality. The truth-maker of an attribution of intentionality to an entity, is the presence of a genuine intentional state in the mind of the entity it's attributed to (1994b, p. 78, p. 82)). For Searle, this will turn out to be a fact about the emergent properties of the brain of that entity (only entities with brains, or made from stuff with the same "causal powers" as brains, could have intentional states).

While Searle thinks that the truth-maker of an ascription of an intentional state would be an emergent property of a person's overall brain, for others (most first-base theorists, as Haugeland (1990) classifies them), the truth-maker will be an actual neurological item or state itself. Appealing to neuroscience here, as potentially providing a more objective truth-maker of an attribution of intentionality to a person, is overly hopeful. The hope is that the intentional state's neural implementation (or neural correlate) can be identified. Thus third parties with brain-scanners could confirm that the intentional state is indeed present in a particular case where it is attributed. In spite of the fact that many clever people think that reductions like this, where we identify an intentional state with a brain state, are a viable possibility,[100] such reductions are impossible. Let me explain why.

One problem with this hopeful move is that first the neural implementation of that type of intentional state would have to be identified. But the only conceivable way to identify the neural implementation of the intentional state would be to *already* have some independent criterion by which to tell when the subject is in fact in that intentional state. It is only if we had that independent criterion, that we could look in (scan) the brain to see which part is active in all

[100] Of course, many people think they are not possible. Token identity theorists argue that particular tokens of mental states have physical descriptions as well as mental descriptions. But most, Davidson (1980) is a prominent example, think that the possibility of specifying the identities, and the laws that correlate psychological events with physical events, is impossible. A particular mental event could have one physical description, and another mental event of the same type could have a quite different physical description. Such theorists, however, would agree with my conclusion –if not with my reasons for that conclusion– that hopes that neuroscience can provide a truth-maker for ascriptions of intentionality are unjustifiably hopeful.

and only cases where that intentional state is present. The intention would be to identify the part or state of the brain that is active when and only when that intentional state is present. But unless we can have some independent criterion for the subject's in fact being in that intentional state, we cannot know that the subject is indeed the same intentional state in each case. The problem is that we cannot have such a criterion.

One reason for this is due to the fact that this would be an inductively generated correlation. The form of the problem here is the same as the form of the problem of misrepresentation, which I talked about back in Chapter Two. We can sensibly ask: is the neurological state associated with just this intentional state and no other? Could it also be associated with some other (not yet tested) intentional states as well? If the brain state is active in all and only the cases *so far* when the subject reports that intentional state (e.g. the intention to do X), this does not rule out the possibility that a similar but distinct intentional state (the intention to do X′) that has not yet been entertained by the subject during testing could also be associated with this brain state. For instance, is the content of the subject's intention to *go wash the dishes*, or to *go wash the dishes unless my favourite movie of all time is on TV*? If we haven't yet tested the brain state in this situation, we cannot make any ruling as to whether this counterfactual situation should be included in the content of the intentional state. This is a problem with any inductively generated law based on a finite set of observations of a correlation. Causal laws are supposed to be counterfactual supporting. But here we have no way of knowing what counterfactual conditions should and should not be governed by the regularity we interpret, and thus we cannot justify claims about the exact content of the generalization. Fodor's "disjunction problem" diagnoses just this problem with generalizations about the content of a particular representational state.[101]

Similarly, we should ask: is the intentional state always associated with this brain state and no other? There may be cases in the future in which this subject reports that they are in that intentional state, but the subject is in a slightly different neurological state. Let's say that one particular Friday, after a hard day at work, the subject claims to intend to do the dishes, but their brain state (B′) is slightly different than the brain state (B) we have so far thought is correlated with that intentional state. Should we look for some distinctive

[101] Compare Fodor's (1990) criticism of Dretske's "Learning Period" idea in "A Theory of Content" (p. 61 ff.).

feature of the subject's present situation, by which we could refine the description of the intentional state we correlate with the brain state? For example, should the correlation be expressed as: when the subject has not had a hard day at work, brain state (B) indicates the intention to do the dishes, otherwise brain state (B') indicates that intention. Or should we refine the description of the brain state, and look for some common features of this brain state and the one we have been correlating with the intentional state, and say that it's this set of common features that is associated with the intentional state? Which part of our theory about the correlation between the intentional state and the brain state should be revised in cases where the correlation breaks down? Is the content of the intentional state in need of refinement, or is the description of the brain state in need of refinement? The answer is at least not *obvious*. We would have to look at *something else* to assist our decision in such cases. The nature of this "something else" is quite telling, and very important for overall argument.

Note that here we are taking the subject's *reports* about their intentional states to be the basis of our correlation. We try to correlate their reported intentional state with the brain state. One obvious problem is that the subject could be lying. But even if we take the subject to be honestly reporting their intentional states, there is a deep problem with attempting to establish a law-like correlation on this basis.

It is assumed that the intentional state is something private, that the person alone has access to. The idea of finding a brain state correlated with this internal state is motivated by a physicalist assumption that states of mind are really states of brains. If it could be shown to be a state of a brain, then we could have a neurological state that someone else (at least someone equipped with brain-scanners) could confirm that the person is in the neurological state they report themselves to be in. We would have a neurological state to use to determine the truth of a person's claims about their intentional states.

The bankruptcy of this notion that *essentially private* intentional states can stand in *any* kind of law-like correlation is shown by Wittgenstein's so-called "private language argument".[102] As Wittgenstein argues using his famous example of the diary-keeper who associates the symbol "S" with a private sensation (§259), there is no criterion for correctness of the association. I cannot

[102] See Wittgenstein's (1958) Philosophical Investigations, § 243 ff.

follow a rule in associating the intentional state with the expression. Because my intentional state (the sensation, for the diary-keeper) is completely private, there can be no difference (not even in principle) between my *correctly* making the association between that intentional state and an expression of our public language, and it merely *seeming to me* that I make the association correctly. We have no criteria by which to tell the difference between these two cases. And this lack of criteria for correctness means that here we cannot speak of being correct or of being incorrect. The whole notion of making the *correct* association would be meaningless; so would the notion of getting the association wrong. But to be able to speak meaningfully of a linguistic expression being correctly correlated with the intentional state itself, we need to be able to speak meaningfully of the *same* intentional state being associated with the expression each time. We need there to be a difference between actually following the rule, and employ the expression correctly, and it merely seeming to me that I employ the term correctly. But there is no criterion by which to tell the difference, if the intentional state is completely private.

Thus if pains, beliefs and intentions are essentially private items, then these private items themselves cannot play any role in public language-games. They are Wittgenstein's "beetle in the box" (§293). Wittgenstein invites us to imagine that everyone has a box with something in it. We call it a "beetle". No one can look into anyone else's box, however.

> But suppose the word "beetle" had a place in these people's language? —If so it would not be used as the name of a thing. The thing in the box has no place in the language game at all; not even as a *something*; for the box might be empty.—No, one can 'divide through' by the thing in the box; it cancels out, whatever it is.
>
> That is to say: if we construe the expression of sensation on the model of 'object and designation' the object drops out of consideration as irrelevant (Wittgenstein 1958, §293).

The words "pain", "belief" and "intention", are used in our language games, but are not used as the names of internal private states. If we can make sense of the idea that people's intentional states can stand in a law-like correlation with public phenomena, then the intentional state cannot be something that is essentially private.

This does not mean that they are the names of something else, however; something that is not essentially private. Expressions for intentional states do not function as the names for *things* at all. Here Wittgenstein is rejecting the bad philosophical picture that people have inner private states of mind, and that these are the referents of our talk about people's intentional states. But he is not replacing it with a different picture where these expression are the names for some other thing, that will be the truth-maker of our uses of such expressions. Wittgenstein's interlocutor (the remarks that Wittgenstein puts in "quotation marks") protests that the intentional state of remembering (say, remembering where I left the car-keys) must be an inner state:

> "But you surely cannot deny that, for example, in remembering, an inner process takes place."—What gives you the impression that we want to deny anything? When one says "Still an inner process does take place here"—one wants to go on: "After all, you *see* it." And it is this inner process that one means by the word "remembering".—The impression that we wanted to deny something arises from our setting our faces against the picture of the 'inner process'. What we deny is that the picture of the inner process gives us the correct idea of the use of the word "to remember". We say that this picture with its ramifications stands in the way of our seeing the use of the word as it is.

The point is that although the picture of intentional states as private inner states, inclines us to think of the truth-makers of attributions of intentional states being the intentional states *themselves*, this is a misleading picture. And it's still that same misleading picture if we take them to be names for intentional states themselves, as things that are not private inner states, but, for example, states of the brain. The inclination to think of such ascriptions of intentional states as referring to intentional states –*things* that are either there or not, and which are the truth-makers for these attributions– is likewise mistaken. Such talk does not play the role of *referring to things* in our language games. (The point I am eventually arguing for, is that such talk functions *normatively*; claiming that someone is in a particular intentional state is rather similar in form to claiming that someone promised to do something.)

As Wittgenstein says (§580), internal (private) states stand in need of outward (public) criteria. The outward criteria of being in the intentional state are the only things that can play a role in public language games that involve talking about our intentional states, and making claims that other people are in certain intentional sates.

This means that in any neurological experiment attempting to identify the neurological correlate of the person being in a particular intentional state, all we can correlate are, not the intentional states themselves, but the outward criteria that we typically use in our language games to justify the use of intentional expressions. What we correlate with the type of brain state we identify, then, are not the intentional states themselves, but the *outward criteria* by virtue of which the person counts as being in the intentional state.

The subject's reports of their intentional states are one thing we could correlate. Putting to one side the concern with the honesty of their reports, as I said earlier, we have reason to be concerned about the specificity of their reports. Is it an intention to wash the dishes, or an intention to wash the dishes or watch TV if a really good film is on? Unless a really good film is on, and the subject knows this, it's hard to identify this as part of the content. The subject might not even acknowledge this caveat until the situation arises.

A deeper reason for this same problem is that, even if the subject is *honestly* reporting their beliefs about their intentional states, we cannot presume that the subject is always *correct*. This kind of first-person authority is a central tenet of Descartes' approach: not even the Evil Demon could deceive me about what I believe. I know the contents of my mind –my desires, beliefs, feelings and intentions– with absolute certainty. Allegedly, I cannot be incorrect about this. However, at least since Freud raised the possibility of unconscious motivations, people have generally accepted the thesis that I can indeed be wrong about my beliefs, desires, feelings and intentions. It is not automatically nonsense, as it would be for Descartes, to for someone to claim that I am wrong about my intentions; that my intentions for doing what I am doing are really something other than what I believe they are, for instance. People's avowals about their intentional states cannot always be taken at face value. The subject's intentional states could be otherwise than what they believe and report them to be.

What is important here, is that in such cases, the claim that I am wrong about my intentions is made, not by appealing to internal brain states, but by appealing to my *behaviour*. I think, and so claim, that I intend to do such and such, but my behaviour is good evidence that I do not really intend this. If you need a truth-maker for my claims about my intentional states, then you need look no further than my behaviour. My behaviour is typically the (defeasible) criterion that people use to justify claims about my intentional states.

To sum up this section: The problem with trying to find a correlation between a type of intentional state and a type of brain state is that we can only form correlations between events or entities that are *observable*. And we cannot take the subject's reports as observations that can be part of a law-like generalization. We can form a correlation only between types of brain state and the types of behaviour (not limited to, but also not necessarily including, what they say) that we ordinarily take as evidence of people's intentional states. Doing so might eventually allow a short-cut: rather than having to looking to what people say and do (something that might require observing minutia of body-language, subtle inflections of speech, and observing over a long period of time), we could look to the brain state correlated with saying and doing that kind of thing. But we wouldn't have captured people's intentional states themselves in any kind of law-like generalization.

By appealing to neuroscience we are left with a problem. For a speech act to be felicitous, the speaker must have certain thoughts, feelings, beliefs and intentions. But if these are private internal states of the person, we are appealing to something that is beyond incorporation into law like generalization, nor into a norm-governed practice. The above argument, however, shows that whatever people's intentional states are, they cannot be private items only accessible to the person who has them. It also reminds us that the truth-makers of people's intentional states are the things that people say and do. It's what people say and do that would be correlated with their brain states.

5.3 *When are we justified in attributing intentional states?*

Daniel Dennett agrees that it's what people do and say that is important here. His take on this problem of what the truth-makers of ascriptions of intentionality are, is a step in the direction I want to go. However, Dennett doesn't quite go as far as I want to go. To Dennett, we need to determine, not whether a system *really has* certain intentional states (what the truth-makers of ascriptions of intentional states are), but whether we would be *justified* in attributing intentional states to the system. The facts that determine whether we are justified in attributing intentional states to the system, to Dennett, are facts about what the system does, not facts about its internal states as first-base theorists like Searle and Fodor maintain. These facts about what the system does, justify attributing intentional states to it, in that doing so enables an observer to successfully predict the system's behaviour. That is, such attributions of intentional states are justified pragmatically.

He gives the example of the "two-bitser" (1995a, p. 404 ff.), a vending machine that is designed to accept US quarters, to argue that we can be *justified* in assigning a function to the system, but that there are no facts beyond functional facts that make it true that this is its function. The two-bitser accepts any US quarters and rejects any other US coins, so we can very successfully predict the system's behaviour (in the US) if we assign to it the function of detecting US quarters; its acceptance state *means* quarter-here-now. It turns out, however, that Panamanian quarter-balboas are identical in size and weight to US quarters and will also be accepted by the device. Dennett uses this example to argue that the device's *function* determines what its acceptance-state means, and the device's function depends on the device's physical constitution, and also on its environment. This is not just a matter of the physical environment, however. Its function is also determined by the intentions of its users and designers. It doesn't just *have* a function, a function is *assigned* to it based on its physical constitution and its environment, but also on what agents do or could *use* it for. This function, then, could be attributed because of what it was designed for, but it can also be based on an engineering analysis of what it would be *good* for. It's function is derived, he says, from human beings' intentions when they use the device and assign a function to it in so using it. The two bitser's function, then, depends on what it is designed for, or on the function it is "exapted" for when, for instance, its quarter-detecting abilities enable it to be pressed into service in Panama as a quarter-balboa detector (p. 406). Its acceptance state, he says, "could mean 'quarter-balboa here now' if we put it in the right environment" (p. 412); that is, if we put it –and used it– in Panama. So, for Dennett, functions like this are attributed to devices like this, depending upon the functions people use them to perform. For Dennett, the function is based on *the intentions of individual agents*: those who design or "exapt" the device to perform that function, and those who use it to perform that function (p. 407). Because a Panamanian soda-pop purchaser can reliably predict that if they feed a quarter-balboa into the machine, it will accept that quarter towards payment for a can of soda-pop, they are justified in attributing to it the function of accepting quarter-balboas. They are pragmatically justified in attributing the function of detecting quarter-balboas to it, and in thinking that that its acceptance state means 'quarter-balboa here now'.

People's internal states, for Dennett, can also have functions, and thus meanings. People attribute functions to such internal states, based on what they are used for. And they are justified in doing so if it attributing this function to the internal state is useful in enabling prediction and explanation of the person's behaviour. These states' meanings, then, are similarly derived from the role they have in producing people's behaviour. "You have internal states, that get their meanings from their functional roles, and where function fails to yield an answer, there is nothing more to inquire about." Thus two different interpretations of what the internal state's function is could each be pragmatically justified, by attending to the behaviour that this internal state brings about. But these two interpretations could disagree, and no further fact could determine which is correct (see Dennett 1991a, p. 49).

Dennett uses Putnam's (1975a) Twin Earth example to argue that the meaning of a person's internal state is derived from the way it is used, that is, by the behaviour it brings about. This meaning can be decided best, he implies, by the individual whose internal states these are (p. 409). Putnam's original example debates whether the internal state that has the function of enabling me to recognize water is correctly applied to the stuff I find when I am whisked in my sleep to Twin Earth (where what they call "water" is a different, though perceptually indistinguishable, substance from what Earthlings call "water"). To Dennett, whether or not it misrepresents the stuff Twin Earthlings call "water" depends on how specific my concept of water is; it depends on the internal mechanism that enables me to recognise water and what it actually is triggered by, and it also depends on the function I assign to this internal state. It depends, for example, on what I would *explain* the concept of water to apply to. I might take the term "water" to mean —and thus take this internal mechanism's function to be to detect— "the stuff we Earthlings call 'water'". However, it could have a looser function for me, such as detecting "the stuff that falls from the sky when it rains and that flows in the rivers and streams, the stuff that people drink, wash with, and swim in." If my concept was this latter concept, rather than the former one, that would make my calling the stuff on Twin Earth that flows from the faucet and rains from the sky "water", a correct application of the concept. It's likely, however, that my concept is rather indeterminate, and it's not clear whether it is correctly applied or not to twin water (Dennett 1995, p. 410). Thus the meaning of a person's water-detecting internal state, for Dennett, depends on the individual and how *that individual* uses that internal state. It depends upon the concept they apply when that internal state is activated and what they take that state's function to be.

Similarly, to take one of Dennett's more Earthly examples, if an expert biologist were to tell you that coyotes are dogs, would you be surprised? If so, then your own DOG concept (and the physical state of you that is responsible for you ability to think about dogs) would be different from that of the expert. Have you previously been strongly wedded either to the view that "by definition", coyotes are dogs, or to the view that they are not dogs? Has the question ever even come up before? Perhaps, he says, your concept has all along had the openness to admit of this new information. Whatever is the case, says Dennett, the meaning of your concept of "dog" depends on the function you assign to it. So for Dennett, this function depends on what things you *as an individual* take the concept (and thus the physical state that implements your ability to recognize and think about dogs) to be correctly applied to.

On Dennett's account, then, for all devices that perform functions, both artificial devices and those that emerge through evolutionary selection ("designed" by Mother Nature), the function is not an intrinsic fact about the system. Rather, intentionality is *attributed to* the system. The system's function is that function attributed to it by an agent who is able to attribute purposes to other entities; that is, an agent able to adopt the intentional stance.[103] And it is attributed by individual intentional agents. And the only justification possible for the particular intentional states that such agents attribute to the system, is whether they subjectively feel that they are better able to predict the system's behaviour because of adopting these intentional stance explanations. They are particularly justified if it is very difficult to account for the system's behaviour from any stance other than the intentional stance. For instance, if the physical stance takes too much computational effort, the "noisy" and more error-prone but easier intentional stance explanations could be well justified.

Furthermore, two different agents could attribute quite different intentional states to the system, that were each incompatible but well justified. They might each enable rather successful prediction of the system's behaviour,

[103] Thus all meaning is derived, says Dennett. There is no "original" intentionality that merits a special status, as the intentionality that other intentionality is derived from. All intentionality is derived from the functions that agents attribute to states of the entities. And this intentionality is part of a network of "mutually supporting" intentionality. You attribute it (intentional states such as beliefs, desires, etc.) to me, I attribute it to you. And importantly, we each also attribute it to ourselves.

and only occasionally make errors. They could disagree about which cases were the successful predictions and which were the errors, though. And there is no fact about the matter, says Dennett, that could settle the question of which one is right (Dennett 1991a, p. 49, see also 1978). Attributions of intentionality, to Dennett, are observer relative, and subject only to justification by that agent, in terms of the pragmatic advantage to that agent in making successful predictions of the system's behaviour.

5.4 *Intentional states are not "hidden" internal states*

Dennett's point is that our intentional states, our thoughts and desires and intentions, are *manifest* in our behaviour. Observers of our behaviour can make very successful predictions of our future behaviour because they can very justifiably attribute intentional states to us based on the way our actions can be interpreted as signs of our intentional states.

The point that our thoughts are not hidden from others, but are manifest in our behavior, is a common theme in Wittgenstein and Ryle. Wittgenstein makes many remarks to this effect.[104] In §573, for example, Wittgenstein asks:

> What, in particular cases, do we regard as the criteria for someone's being of such and such an opinion? When do we say: he reached his opinion at that time? When: he has altered his opinion? And so on.

Wittgenstein goes on to answer this question, not by pointing to inner processes, but by presenting a range of examples of cases where people's intentional states are manifest in their behaviour. In *The Blue and Brown Books* (1958/1933-6, p. 20) Wittgenstein explicitly recommends this tactic of giving a large series of examples as a cure for the idea that there is an inner essence to intentional states such as wishing, thinking, meaning, understanding, and so on.

> If we study the grammar, say, of the words "wishing", "thinking", "understanding", "meaning", we shall not be dissatisfied when we have described various cases of wishing, thinking etc. If someone said "surely this is not all that one calls "wishing', " we should answer, "certainly not, but you can build up more complicated cases if you like." (Wittgenstein 1958/1933-6, p. 19)

[104] See for example, Wittgenstein (1958) §316-367, esp. §330-331; also §572-589; p. 223.

For instance, in §576 he says: "I watch a slow match burning, in high excitement, follow the progress of the burning and its approach to the explosive." He says this description of what he *does* is "certainly a case of expecting". He also points out that this is a case of expecting, whatever he may be thinking. In this case he "might not think anything at all or have a multitude of disconnected thoughts." He gives similar examples of expecting someone, and of hoping that he'll come (§577, §584-6), where he describes the behaviour that is a manifestation of the hope and the expectation. Expecting him to come gets a thorough treatment in *The Blue and Brown Books* (1958/1933-6, p. 18):

> What happens if from 4 till 4:30 A expects B to come to his room? In one sense in which the phrase "to expect something from 4 till 4:30" is used it certainly does not refer to one process or state of mind going on throughout that interval, but to a great many different activities and states of mind. If for instance I expect B to come to tea, what happens *may* be this: At four o'clock I look at my diary and see the name "B" against today's date; I prepare tea for two; I think for a moment "Does B smoke" and put out cigarettes; towards 4:30 I begin to feel impatient; I imagine B as he will look when he comes into my room. All this is called "expecting B from 4 till 4:30".

In this description, the "states of mind" that Wittgenstein refers to themselves are exhibited in his behaviour. He thinks for a moment "does B smoke?" and then sets out cigarettes. His feeling impatient also is exhibited in his behaviour; say looking often at the clock and the door, pacing and fidgeting, having difficulty concentrating on any other task, being roused by any noise that might be a sign of B's approach, and so on. §579 of the *PI* consists of the single question: "The feeling of confidence. How is this manifested in behaviour?" Similarly for the intentional state of believing Goldbach's theorem:

> "... Let us look and see what are the consequences of this belief, where it takes us. 'It makes me search for a proof of the proposition.' —Very well, let us look and see what your searching really consists in. Then we shall know what the belief in the proposition amounts to" (§578)

In §587, Wittgenstein points out the problem with relying on introspection to tell what I believe. Even when we do this, he says, we determine what we believe, intend, feel, and so on, by imagining how these intentional states would manifest themselves in our behaviour:

> It makes sense to ask: "Do I really love her, or am I just pretending to myself?" and the process of introspection is the calling up of memories; of imagined possible situations, and of the feelings that one would have if....

These manifestations of our intentional states are often linguistic, as Wittgenstein points out in §585. What I am inclined to say, as much as what I am inclined to do is a manifestation of my intentional states:

> 585. When someone says "I hope he'll come" —is this a *report* about his state of mind, or a manifestation of his hope? —I can, for example, say it to myself. And surely I am not giving myself a report. It may be a sigh; but it need not. If I tell someone "I can't keep my mind on my work today: I keep thinking of his coming"—*this* will be called a description of my state of mind.

Thus to Wittgenstein, what we call "states of mind" are in fact states of a whole person.

> 573. To have an opinion is a state.—A state of what? Of the soul? Of the mind? Well, of what object does one say that it has an opinion? Of Mr. N.N. for example. And that is the correct answer.

These intentional states are manifest in what the whole person does.

Furthermore, the *criteria* for being in that intentional state also involve what the whole person does. Norman Malcolm summarizes this line of thought in his "Thinking" (1978), where he argues that what we typically call "inner" processes and abilities must be manifest in peoples' activities.:

> the ability to multiply in one's head logically presupposes the ability to multiply aloud or in writing. And that is *necessarily* so. For a person who was *not able* to execute any of the processes of multiplication, aloud, in writing, or in other outward signs, could not be said to be multiplying in his head, even if he was usually able to produce the right answer to multiplication problems!...
>
> Isn't there some similarity here between calculating and thinking? ...Thinking in one's mind (silent thinking, pausing to think) is not the most fundamental form of thinking, but instead presupposes thinking in play, work, or words" (p. 415).

If the idea that the person could produce the correct answers, but couldn't execute any of the processes of multiplication aloud or in writing, seems absurd to you, imagine that the person has a calculator hidden in their pocket and is secretly employing that device to arrive at the answers. Such a person could not be said to know how to multiply numbers in his head, even though he produces the right answers.

Similar thoughts can be found in Ryle's *The Concept of Mind* (1949). Here Ryle points out that having certain thoughts and feelings amounts to having abilities and dispositions to behave in certain ways. "Dispositional verbs," he says (p. 114), "like 'know', 'believe', 'aspire', 'clever' and 'humorous'... signify abilities, tendencies or pronenesses to do, not things of one unique kind, but things of lots of different kinds". This is Ryle's overall point: that "X has a mind" means not that X has a special place where private thoughts happen. Rather it means that X has certain abilities, tendencies and pronenesses to *do things*. "X has a mind" is properly predicated of entities that are capable of answering thoughtfully, of adding numbers carefully, of standing to attention obediently, and so on. In general, it is used if X is capable of acting intelligently.[105]

Thus on Ryle's view, saying of a non-human system, as Searle (1990) does, "It doesn't really have intentional states, although it behaves in every way indistinguishable from the behaviour of something that does," is infelicitous. It's infelicitous in the same way that "I have locked him up in the room; there's only one door left open" would be.[106] If the second clause is true, then the speaker is not following the normal procedure for using the first expression to assert something. To Ryle, having a mind (having intentional states correctly predicated of one) just is being able to behave, and actually behaving, in the ways that en-minded beings behave. Thinking is not some "extra" process going on "behind" the behaving. To Ryle, our thoughts, beliefs, desires and intentions all are expressed or exhibited or manifest in our behaviour —especially in our linguistic behaviour.

In Chapter Two I discussed similar remarks made by Peirce "A Survey of Pragmaticism" (5.476 ff.) Peirce makes similar points about how a thought or a concept must be exhibited in behaviour. Recall that for Peirce, a sign is the kind of sign it is by virtue of the interpretant generated in an interpreter. This interpretant is usually a thought or concept. Such mental signs, however, pose a problem for this view. The interpretant of a thought or "mental sign" cannot

[105] See Bestor (1979) for an argument that this is Ryle's overall point, and for Ryle's last words on what he was trying to achieve in The Concept of Mind in which Ryle confirms this description.

[106] C.f. Wittgenstein (1958) §99.

always be another mental sign, due to infinite regress. The regress has to halt somewhere at an interpretant that is not itself a sign (and so doesn't have an interpretant). Thus the *"ultimate* logical interpretant of the concept" is what Peirce calls a *habit change,* "...meaning a change in the person's tendencies toward action..." (5.476). Being in a particular intentional state, to Peirce, must at some level be manifest in the way a person acts, or the way they are (counterfactually) prepared to act should certain circumstances arise.

Sometimes, however, we can keep our thoughts and feelings and intentions to ourselves, but this skill is acquired, and requires some effort:

> This trick of talking to oneself in silence is acquired neither quickly nor without effort; and it is a necessary condition of our acquiring it that we should have previously learned to talk intelligently aloud and have heard and understood other people doing so.... People... tend to suppose that there is a special mystery about how we publish our thoughts instead of realizing that we employ a special artifice to keep them to ourselves." (p. 28).

The common theme for these theorists, one with which I strongly agree, is that in most cases our thoughts are not hidden and private, but are exhibited in the things we do, and in the subtle ways in which we do them. It takes a large amount of effort, and lots of practice, to be able to hide our thoughts, feelings and intentions from others; as any poker player who has either tried to bluff, or tried to hide their excitement at having a winning hand, could attest.

In every case of attributing an intentional state to someone, then, the final criterion for the felicity of the attribution is the person's verbal and non-verbal behaviour. We have nothing else to appeal to. But often we need nothing else to appeal to.

5.5 "Mental" actions are activities of whole persons

A complement to the view that people's behaviour manifests their thoughts, feelings, beliefs and intentions, is that there is no longer a distinction to be made between "mental" behaviour and "bodily" behaviour. Contrary to many theorists' intuitions, I use the term "behaviour" to refer to people's "mental" activities as well as to their "bodily" activities. Once we accept the view that the fundamental unit of analysis is a whole person and their activities, and not a dualistic entity with a thinking part and an acting part, then we also accept the view that so-called "mental" activities are just as much things that whole people do as more bodily, moving things about in the world kind of activities. Here so-called "mental" actions like thinking about something, wondering about its location, trying to remember where I put it, worrying that I won't find it, and remembering where I left it, are just as much things that I do, as picking the object up and putting it in my pocket. In fact, the distinction between the two breaks down if we think about the way many of our mental activities are carried out: adding up numbers using a pencil and paper, for instance, is an activity happening in a complex interactive system, involving the pencil and paper, and also my hands, eyes and central nervous system. As Wittgenstein points out, "expecting him to come to tea" is not a mental activity, but the activity of a whole person. It is manifest in all the things that Wittgenstein does during the period of time that "expecting him to come to tea" is felicitously predicated of him: making tea, setting out cigarettes, impatiently pacing and checking the time, attending expectantly to noises outside the door, and so on.

The only principled difference between thinking about an object, and picking the object up is that the action is more "publicly observable" in the latter case. Thinking about making coffee is as much something a person *does* as making coffee.[107] Each of these is something a whole person does. Some of these things people do are more "publicly observable", but this is a difference in degree, not a difference in kind. (This point is brought out more in the next section)

5.6 *The normative structure of attributions of intentionality*

Although Dennett (like Ryle and Wittgenstein) incorporates this position that our intentional states are manifest in our behaviour, he misses an important part of the *reason why* our intentional states are so manifest. Ryle also seems to miss this reason. Wittgenstein recognizes it to some extent.[108] Brandom (1994) emphasizes it. To Dennett, the evidence on the basis of which an intentional system attributes an intentional state to another entity, is the entity's behaviour.

[107] A comment by Wittgenstein (1958, p. 190) relates here: "Don't look at it as a matter of course, but as a most remarkable thing, that the verbs 'believe', 'wish', 'will' display all the inflexions possessed by 'cut', 'chew', 'run'."

[108] It's unclear to what extent Ryle recognized this. Wittgenstein recognizes it to some extent; but to what extent would be a matter interpretive dispute, I'm sure. An interesting tangent that I haven't space nor time to explore here, would be to try to amass textual support for a position on this question.

Dennett's standards of justification, the standards that the evidence must meet, however, are very much the subjective standards of the individual agent that attributes the intentional state, and whether that agent feels (pragmatically) justified. Dennett takes all intentionality to be derived from the functions that individual agents attribute to systems. This is partly why Dennett calls himself a "milder than mild relativist" (1991a). Intentionality is attributed relative to the conceptual schemes and pragmatic goals of individual agents. Dennett's examples (e.g. 1995a, Chapter 14) make it clear that the standards of justification are standards internal to individual agents. Whether the agent would be justified in attributing intentionality to another entity, depends on whether it would be rational for the agent to attribute the intentions.

For me, intentionality is also attributed. However, it is attributed relative to *shared* standards of justification: the norms at play in the practices of a community, rather than to individuals' conceptual schemes. It's these shared standards, accepted both by attributers of intentionality and by the people it's attributed to, that ensure that our intentional states are manifest in our behaviour.

Austin's *felicity conditions* and Wittgenstein's talk about the *grammar* of certain "mental state" expressions gets us some way to seeing the normative practices within which attributions of intentional states are attributed. These norms link people's behaviour to what we normally say about people's intentional states.

For Wittgenstein the term "grammar" covers more than just the syntax of the term, as it does for linguists such as Chomsky. For Wittgenstein "grammar" is used to refer to the way the word is used, it refers to the role that the term plays in people's language-games, and in the form of life those games constitute. [109] For example, the grammar of the word "belief" is the role that the word "belief" and its variants (e.g. "believes") play in our shared form of life. When Wittgenstein discusses the grammar of folk-psychological terms, he relates the attribution of particular intentional states to the network of norm-governed social practices (forms of life) within which the practice of making such attributions is embedded. The following collection of remarks bring this out well:

[109] Easton (1978) also bases her argument on claims about the grammar of expressions, citing Wittgenstein's position that the grammar of an expression is given by the role the expression plays in people's lives, practices, and conventions.

> §370. One ought to ask, not what images are or what happens when one imagines anything, but how the word "imagination" is used. But that does not mean that I want to talk only about words. For the question as to the nature of the imagination is as much about the word "imagination" as my question is. And I am only saying that this question is not to be decided—neither for the person who does the imagining, nor for anyone else—by pointing; nor yet by a description of any process. The first question also asks for a word to be explained; but it makes us expect a wrong kind of answer.[110]

Similarly, to Wittgenstein, we should not ask what beliefs are, but ask how the word "belief" is appropriately used. This question is as much about the nature of belief as the original question. But instead of encouraging us to expect an answer in terms of what kind of thing (e.g. what kind of neurological state) a belief is, we get an answer in terms of the conditions under which it is appropriate to use the term "belief", and our agreement in judgements about when the term appropriately applies. In what situations is it appropriate to use the term, to whom, by whom, in what situations and contexts, to perform what speech acts, etc.? This question is to be answered by appeal to our shared criteria for when a situation licenses using the word "belief" to describe it. That is, we settle it by appeal to our "agreement in judgements" (§242) about the cases, such as whether, for instance, "Mason believes that he can get the chapter finished this evening" is appropriate to use. Wittgenstein's answer, as it was for the state of "expecting that he will come" is to look at a wide variety of cases where we attribute the mental state. But it's not to look for some common ingredient to all these cases (certainly not a state of an actual mind nor the state of a brain. Instead he points out the cases and the fact that we do indeed agree that these count as cases of believing that the chapter will be done, or intending to get the chapter done. In the cases where we disagree, there is still room to make arguments for one side or the other. And what we appeal to in such arguments will be the *public* circumstances and details of the particular case. The agreement in form of life, is agreement in judgements about whether a public, shared, norm

[110] It may help to see this in the context of a few other remarks from the PI:
§371. Essence is expressed by grammar.
§373. Grammar tells what kind of object anything is. (Theology as grammar.)
§241. "So you are saying that human agreement decides what is true and what is false?"—It is what human beings say that is true and false; and they agree in the language they use. That is not agreement in opinions but in form of life.
§242. If language is to be a means of communication there must be agreement, not only in definitions but (queer as this may sound) in judgements....

is followed when we attribute a particular intentional state to an agent, based on details of what the agent does.

The speech acts people perform and the criteria by which we judge their felicity, attempt to explicitly highlight some of the (mostly tacit) norms by which we (as members of a common form of life) make such judgements. That these norms are shared norms, is the reason that for the most part we agree in judgements. And where we disagree in judgements, dissenters will usually be pointing to failures to satisfy the kinds of criteria that the felicity conditions lay out. Take, for instance, the felicity conditions on your attributing to Mason the intention to get the chapter done tonight. Under what circumstances would such an attribution be made appropriately? If you attribute to me the intention to get this chapter written tonight, what felicity conditions would this act have to meet? Or more to the point, what *would I have to do in order for this intentional state to be felicitously attributed to me?* What circumstances would make such an attribution infelicitous? The details are not difficult to fill in. A couple of examples should suffice to point out the normative role that the felicity conditions on attributing this intentional state to me.

For your act of asserting "Mason intends to get the chapter written tonight" to be felicitous, then, for example, I must have behaved in ways that should –according to the shared criteria (norms) for the appropriate use of this expression– lead one to attribute that state to me. (A.2.1: *The circumstances for performing the act must be appropriate.*) I could have declared this intention, for instance. I could alternatively have refused an invitation to do something else this evening, have declared that I am close to finishing, have stocked my office with supplies, informed my partner that I will be home very late, disconnected the email and phone, and have stayed in my office working diligently apart from to get coffee or to dispose of the coffee after it's run its course through my system. Doing all this would license that your attribution of this intentional state. If, however, you knew that I accepted that invitation, declaring that I will finish it in the morning, then this attribution would not be felicitous. It would be as infelicitous as "I have locked him up in the room. There's only one door left open" would be infelicitous. Someone asserting this, in such a situation, is not using the expression in the way that the norms for using it stipulate it should be used. (Compare Austin's conditions A.1.1: *There must be a customary procedure for performing the act…* and B.1: *the procedure must be executed correctly.*)

My future behaviour is also relevant: if I did everything I just listed, but then decided to stop writing and go home to watch TV, then although you might have evidence for asserting this, it would be declared infelicitous because my after-the-fact behaviour raises suspicions, at least, that I didn't really so intend (I was pretending for instance).

This raises an different felicity condition: In order for the attribution to be felicitous, it's not just that I must have behaved in the ways that license attribution of the intention to me, you must believe that I have so acted (C.1.1). You must believe that I do intend this. Otherwise you are speculating, not asserting that I intend to finish the chapter tonight. For instance, you must also have seen or heard me doing the things listed above. You must be in possession of some evidence for declaring this. You must, in other words, be the appropriate person to make such an attribution (A.2.2). By asserting this, furthermore, you take on a certain obligation to do certain things (C.1.3 and C.1.4). For instance, if challenged, you should recognise an obligation to explain your grounds for asserting this. (Although some other intentional states may be so nebulous that you cannot state explicitly what leads you to declare this: e.g. "He really doesn't love her, he's just pretending to himself.")

The principal felicity condition I want to draw your attention to right now (other conditions will be highlighted soon) is the first one: A.2.1: The circumstances must be appropriate. The "appropriateness" of the circumstances, involves my past present and future behaviour, including my speech acts. This "appropriateness" is a normative term. You should not assert this unless the circumstances are appropriate. What those precise circumstances have to be to be "appropriate" are rather nebulous. The point, however, that Wittgenstein raises, is that we *can* tell when circumstances are appropriate.

My point here is that whether I am justified in attributing a particular intentional state to an entity, depends not on individual agents' dispositions to attribute intentionality as Dennett argues, but on a community's *shared* standards of whether the speech act of attributing a particular intentional state to that system would be felicitous. We determine whether an attribution of a mental state is felicitous by appeal to the social and normative roles, as well as the pragmatic roles, that attributions of intentional states play in people's shared forms of life. Attributions of intentional states are not justified by virtue of whether the person's mind really contains that mental state (Searle, Fodor) or whether the person's brain really is in a certain neurological state (Searle), nor by whether the individual attributing the intentional state feels (pragmatically) justified in making the attribution (Dennett). Rather, they are justified by virtue of whether the person's behaviour, and the social and physical context in which it is embedded, accords with the community's shared criteria for attributing that intentional state. The criteria that we use to determine the felicity (or appropriateness or social okay-ness) of attributing certain intentional states to others are *public* criteria—criteria involving the ways people behave and speak.

5.7 *Attributing intentional states to oneself is also subject to public norms*

It's important to emphasize, again, that my uses of terms of folk psychology to attribute intentional states to myself are also subject to public norms, and assessable for their felicity in terms of public criteria. They are just as subject to publicly shared norms as are attributions of intentional states to other people. When you attribute an intentional state to me, you do so because the norms of the practice of attributing intentional states as reasons for people's actions entitle you to infer from my actions that I am in that intentional state. I do the same thing for myself. I do not *know* that I am in particular intentional states (that would be an infelicitous use of the word "know"). I do not necessarily *infer* that I am in them either, although this might be the case in some circumstances. Generally, I tacitly follow the same norms of the practice of attributing intentional states by virtue of the way they behave, that I follow when I attribute them to other people. I *attribute* intentional states to myself; the intentional states one should attribute to me based on my behaviour (and on my dispositions to behave in counterfactual situations). I simply *hold* or *take* myself to have those intentional states.

As Ryle (1949) points out, "The sorts of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same" (p. 149). The major differences, Ryle says, are "differences in the supplies of residual data". They are not differences in the type of data the attributions are based upon.

> I learn that a pupil of mine is lazy, ambitious and witty, by following his work, noticing his excuses, listening to his conversations, and comparing his performances with those of others. Nor does it make any important difference if I myself happen to be that pupil. I can indeed then listen to more of his conversations, as I am the addressee of his unspoken soliloquies; I notice more of his excuses, as I am never

> absent, when they are made. On the other hand my comparison of his performances with those of others is more difficult, [not because their performances are "hidden" from me, but] since the examiner himself is taking the examination, which makes neutrality hard to preserve and precludes the demeanor of the candidate, when under interrogation, from being in good view (p. 162).

The difference between attributing intentional states to myself, and attributing them to others, is largely due to a difference in perspective, than a difference in the kind of information I have to go on. I have better access to "data" on which I base attributions of intentional states, so my attributions are more informed. But I also have more at stake here. My intentional states are part of my self-image, and it matters to me what kind of person I am.[111] For instance, because it matters to me that I am the kind of person, say, with honorable intentions, I might attribute such intentions to myself in a certain situation, in spite of the fact that someone with less at stake might justifiably attribute less honorable intentions to me.

There are certain attributions of intentional states, (in certain circumstances) that the person they are attributed to is best qualified to make. (Notice felicity condition A.2.2: The persons performing the act must be appropriate.) Thus I am perhaps best qualified, for example, to attribute to myself the intention to be home in time for supper. However, in some circumstances I might only attribute particular intentional states to myself after examining how I have been behaving. This behaviour, of course, includes my silent soliloquies. It also includes how I am inclined to act in various counterfactual situations: my declarations (to myself) about how I would feel, react, what I would say, etc. in various counterfactual situations, are also part of the "silent soliloquies" on which I base the intentional states I take myself to have.

Sometimes, in fact, others' claims about what I believe are easier to justify than my own claims. People generally accept, at least since Freud, that in certain circumstances someone could argue that my claims about my beliefs and desires are incorrect (infelicitous). And such arguments will be defended by appeal to my behaviour and other public matters. It's certainly not *automatically* infelicitous to tell someone, "You claim that you believe this, but look at how you've been

[111] Alisdair MacIntyre's *After Virtue* (1981), discusses similar themes, about the narratives we tell ourselves, and the role that the norms that go with being that kind of person shape the motives we feel, the commitments we undertake, and so on.

acting! It's clear to me, even if it isn't clear to you, that you don't believe that at all."[112]

5.8 *Felicitously attributed intentional states entail obligations.*

I said earlier that there is a deeper reason why people's intentional states are manifest in their behaviour. As the developmental psychology literature I discussed in Section 4.4 indicates, we human beings quite naturally view one another's action as purposeful and attribute intentional states as reasons for the things we see people do. Doing so is a very useful way of predicting their behaviour. And as I have been arguing, part of the reason for this predictive success is that attributions of intentionality are governed by norms that stipulate what intentional states it is appropriate to attribute to someone based on their behaviour.

However —and this is the important part— the norms on this practice also govern *what someone who is in a certain intentional state should do.* Attributing an intentional state to someone licenses certain expectations of that person. Thus the norms do not just guide expectations of future behaviour just in a predictive sense, but a normative sense. There are certain things I *ought* to do, and certain things I *ought* not do, if I am in that intentional state. That's the normative force of "rational behaviour" speaking there: It would be rational for me to do such and such if I believe so and so, and of course *I should behave rationally.* The desire to behave rationally, to be seen to be a rational member of one's community, has a peculiar force on human beings. As I'll argue in Chapter Six, this is no small matter; there are good evolutionary reasons for instilling a desire to conform in certain social species.

It's this normative practice, and the normative force of "rational behaviour" that underlies the predictive successes that Dennett points towards. He attributes the predictive success of attributing intentional states to others simply to a well-chosen *explanatory* and *predictive* system. However, Dennett ignores the fact that it is also a *normative* system. It's a normative system in two senses. This practice of attributing intentional states based on their behaviour to others is shared, and being shared, it also constrains the behaviour of the people the intentional states are attributed to. These constraints come from the fact that others are *entitled* to expect certain behaviour of me, if particular intentional

[112] C.f. Ryle (1949), pp. 155-6.

states are felicitously attributed to me. Just as I am committed to performing certain actions if I do indeed intend to get this chapter written this evening, so others are entitled to expect me to behave in that way.

Thus if someone attributes to me the intention to get Chapter Five written by the end of the evening, and I agree that this is felicitously attributed to me, then I undertake an *obligation* to conduct myself in a certain way in the future (Austin's Condition C.1.3). Brandom puts this similarly in terms of my being *committed* to behaving in a certain way. Brandom sees the normative practice of giving and asking for reasons as being played by *deontic scorekeepers* (Brandom 1994, pp. 157-66). He uses this term to indicate the way that, within the practice of explaining behaviour in terms of reasons, we keep score on one another. Attributing intentional states to others is a way of keeping track of the commitments we are entitled to expect them to live up to, and to exhibit in their behaviour. If the intention to get Chapter Five written tonight is felicitously attributed to me (and if I recognize that it is), then I am committed to try to get it done. So it's very easy to predict what would happen if a friend called and invited me out to a movie. I would say that I can't do that *because* I'm trying to get this chapter written by the end of the evening. The intentional state is given as the reason for my declining the invitation, and this is accepted as a valid reason because both myself and the person inviting me out recognize the commitments that go along with being in that intentional state.

These norms, and the practice of treating people as being committed to various types of behaviour based on the intentional states that are felicitously attributed to them, also save Austin's felicity conditions, particularly the felicity conditions on attributing intentional states to others from a potential regress problem. Earlier, when listing the ways that attributing to me the intention to get Chapter Five finished by the end of the evening could be judged infelicitous, I pointed to condition C.1.1: the participants must have the requisite beliefs. Thus one of the felicity conditions on your attributing this intention to me, is that you believe that I have this intention. But how do we determine if this condition holds? What are the felicity conditions on my attributing to you the belief that I have that intention? One such condition will be that I believe that you have this belief that I have that intention. And so it goes; the regress appears unavoidable.

There are two solutions to this regress problem, both based in the normative structure within which intentional states are attributed. Brandom points out one solution. He argues that an interpretation of something as having intentionality, cannot always itself be an intentional state. That intentional state is itself in need of interpretation to determine its content. Brandom's (1994, p. 61) solution to the regress of interpretations is similar to Peirce's: ground it in behaviour, in something that is not itself in need of intentional interpretation. For Brandom, we move from explicit, contentful attributions of intentionality to others, to implicit *practical treatment* of those assessments as correct or appropriate: we sanction certain assessments as incorrect, and reinforce certain others. This is Wittgenstein's "agreement in judgements" about whether a rule has been followed.

The norms of our practices, as Heidegger argues (see Chapter Three), are to a large extent non-explicit. A large amount of norm-guided behaviour is brought about not by explicitly consulting the norm, but by conforming to the norm nonetheless. Wittgenstein famously raises the problem with *interpreting* the norm, such that we can explicitly state what the creatures ought to do, and can determine precisely which cases count as violations of the norm. The problem with interpreting the norm, is that any number of possible interpretations can be constructed that only differ in the counterfactual cases. Wittgenstein's solution, which Brandom elaborates well, is to look not at interpretations of the rule, but at the *practice* of taking or treating certain actions as correct and others as incorrect:

> ... What this shews is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases (§201)

For example, behavioural conditioning will bring it about that a creature follows the norm it is trained to follow, while it has no explicit representation of the norm. "Following a rule" says Wittgenstein (1958, §206), "is analogous to obeying an order. We are trained to do so; we react to the order in a particular way."

Much of our attributions of intentionality are not explicit in beliefs about one another's beliefs, but implicit in the ways we treat the person. We treat people as our norm-governed form of life "trains" us to treat them: we treat them as "we" in fact do treat someone with those beliefs and intentions, without necessarily explicitly thinking that they have these beliefs and intentions. Human beings *can* explicitly think about those beliefs and intentions, but we don't *have* to. We can just follow the normative guidelines for treating people as though they have certain beliefs and intentions, and simply take such attributions of intentionality as correct, by responding to them in the appropriate ways. As Brandom argues, this practical way of grasping a norm, by "taking or treating performances as correct or incorrect, by responding to them in practice" (p. 63) is at the foundation of all normativity. The foundation is not interpretations about the content of norms, and justifications for interpreting the rule in this or that way. The ground is the way we just treat certain acts of assessment as correct, and agree in judgements about which ones are correct and which incorrect. Wittgenstein again, from *On Certainty*:

> Giving grounds, however, justifying the evidence comes to an end; —but the end is not certain propositions striking us immediately as true, i.e. it is not a kind of *seeing* on our part; it is our *acting*, which lies at the bottom of the language-game. (Wittgenstein 1969, §204)

And again, from the *PI*:

> 217. "How am I able to obey a rule?"—if this is not a question about causes, then it is about the justification for my following a rule in the way I do.
>
> If I have exhausted the justifications I have reached bedrock, and my spade is turned. Then I am inclined to say: "This is simply what I do".

Wittgenstein's ground for the rules of a language game, a *shared* norm-governed practice, however, is not that this is what *I* do, but that this is what *we* do. Without a ground to our norms in our *shared practice* of treating certain actions as correct or incorrect, all normativity is open to the problem of a regress of interpretations. Brandom (1994) concludes, rightly, that without a ground in the practice of simply treating certain actions as appropriate and certain others as inappropriate, "No sense could be made... of the difference between acting according to the rule and going against it" (p. 21).

Thus you do not need to explicitly believe that I have this intention. Often your treating the attribution of intentionality as felicitous is not because you believe that I have this intention, but that you treat me as one would treat someone with that intention. You expect certain behaviours of me, and are surprised what I don't behave that way. For instance, if you find out the next morning that the chapter still is not finished, you feel justified in asking for an explanation for why I failed to finish the chapter last night.

My behaviour toward you in this situation, in treating you as entitled to an explanation, points towards the other solution to this potential regress problem, which I deal with in the next section.

5.9 *Attributing intentional states to oneself is endorsing the obligation*

The attribution of intentionality allows you to predict and expect certain behaviours of mine, in response to certain situations. But when I attribute intentional states to myself, I do not just do so to *predict* my own behaviour. I use it to constrain my behaviour. I do not just expect that I will behave in certain ways, I undertake an *obligation* to behave that way. In this sense, attributing an intention to myself is the equivalent of making a promise to myself (it's the equivalent of making a promise to you, if I give you reason to attribute that intentional state to me too). It becomes part of my self-conception that I am a person with such and so beliefs, desires, goals, and so on. Exhibiting the behaviours that go with that self-conception is part of maintaining that picture of who I am.

And here there is no regress problem. When attributing intentional states to myself, I do not believe that I am in that intentional state. I certainly do not believe that I believe that I am in that intentional state. I neither believe nor know that I am in that intentional state. I simply treat myself as being in that intentional state.

This is where the normative force of "rational behaviour" really has some force. It's also where the norms relating behaviour to the intentional states that are reasons for the behaviour, run into their mirror-image in the norms licensing expectations of behaviour based on intentional states. Treating myself as being in a particular intentional state, is to hold myself to be under the obligation to behave as the norms relating intentional states to types of actions would lead one to expect that I should. I endorse the obligations others expect of me, by (even tacitly) agreeing that their attributions of intentional states to me are felicitous.

But people don't usually go around declaring to one another the intentional states they take each other to be in. This does happen, but as I argued in the previous section, most of the time we just *treat* one another as being in those intentional states. This means, then, that acting in ways that would license attributions of intentional states, is also *tacitly endorsing* the obligations that come with those intentional states. If I behave around you in ways that license your attributing a certain intentional state to me, then it is expected that I endorse the obligations that come with being in that intentional state.

By behaving in ways that license these intentional ascriptions, then, I license people around me to hold me to the obligations that go with those intentional states. I license them to expect me to behave in the ways someone in that state should (*norm*ally) act. These obligations, and the fact that others will be justified in expect me to honor them, are the force behind the norm that I *should* act rationally. Promising to do something, then, is simply making it explicit that my acknowledgement of the obligation that I would tacitly endorse by simply acting in ways that license people take me as intending to do it.

5.10 Derived and original intentionality, revisited.

I have been arguing that the normative practice of attributing intentional states to people (others and ourselves) as reasons for actions, and the practice of taking people to be committed to performing certain types of actions based on the intentional states so attributed, undergirds and supports all intentionality. Searle argues that one cannot have derived intentionality without something having original intentionality for it to be derived from. So: is there anything according to my theory that has original intentionality?

Intentionality, as I see it is institutional. It is a feature of certain practices that certain actions within them count as intentional, just as it's a feature of certain tennis-playing practices that certain actions within them count as serves, fouls, and match-winning aces. Thus all intentionality is instituted by, and so we might say derived from, the practices in which we participate, and the norms that govern those practices. I am hesitant to call it derived, however, simply because using this word implies that there are some things with original intentionality, for this intentionality to be derived from. These practices and norms do not have intentionality. Nothing has original intentionality, in Searle's sense. There is no such thing as original intentionality, if by this we mean a property that certain things have by virtue of which they *just are* about or directed towards other things.

If it is all derived, there are still distinctions to make, however. There are certainly levels of abstraction to deal with here, and very "primal" types of norms and actions and practices out of which more explicit and conscious actions and norms and practices arise. Within these higher-level practices, intentionality arises and is instituted.

The "background" that Heidegger, Wittgenstein, Searle and Dreyfus talk about, is certainly a good place to start. There are actions, practices and norms aplenty at this level. All of them are completely tacit and non-conscious and non-intentional. Very simple creatures have their entire form of life at this level; much of human beings' forms of life is also based in this type of actions practices and norms as well. (This tacit level is the condition for the possibility of explicit, contentful, conscious intentionality). At this very primitive level there are purely causal mechanisms, that are often not cognitively penetrable, and often "hard-wired" (especially in non-humans), either by the process of natural selection, or by behavioural conditioning. Here we find norm-abiding behaviours that have been trained into the creatures, mostly by their conspecifics treating certain behaviours as inappropriate and appropriate, by sanctioning or rewarding them (doing something that increases or decreases the chance of their recurring).

The mechanisms that enable such behaviour to take place are causal mechanisms, but they do have some right to be described as "intentional". However, if we, as intentional systems, look at this level we cannot justify ascriptions of intentional content. Any number of incompatible interpretations (that disagree only in counterfactual cases) could be made of the rule that constrains the behaviour. And the rule that the interpretation is based on itself stands in need of interpretation. And so the regress goes. An explicit attribution of content here by an intentional system, therefore, is presumptuous. If we attribute functions or meanings to the causal mechanisms, they will be indexical signs at best. Certain actions can (for us observers) count as being directed at certain objects. There is no room for regularities like this to have content, the systems cannot misrepresent (that's why the disjunction problem, applied to frogs is intractable). Here the system's actions count as being directed at *whatever* objects cause the activation of the causal mechanisms that result in the action: the frog snaps at *that*; I blink as *that* comes towards my eye. But there's no content at this level.

At higher levels —built on top of this "background" of tacit skills, practices and norms—, we have the possibility of the consultation of explicit norms. We can do as we should do, and do it because that's what we should do. We are not *caused* to follow the norms, but can *choose* whether or not to do as we should. The norms themselves exert a force on us, however, the force of the obligations our behaviour entails. Here attributions of content are possible (although not always necessary). Here content is attributed to whole people, based on their actions; to explain their actions; as *reasons* for their actions.

According to the normative practices, certain *actions* count as being directed at particular objects and states of affairs. They count as being directed at these objects, by virtue of the norms that stipulate the reasons it would be appropriate to attribute to the agent. Here we can assign contents to those intentional states. We can even assign contents (intentions) to actions that relate the action performed to abstract, absent, and counterfactual objects and states of affairs the action counts as being directed towards. We can assign *symbolic* contents. The actions count as being directed at objects, *as objects of a particular kind*. The intentionality of our actions, is instituted by the practice of attributing reasons for actions. The directedness and content of the action is by virtue of the directedness and content of the intentional states that count as being the reason for that action. And these intentional states are instituted by the norms of the shared practice of giving and asking for reasons.

And what of the internal mechanisms and states that enable the agent to perform these actions? What of the internal "representations" that enable such agents to perform such "representation-hungry" actions? All the intentionality I have been talking about so far is derived, in the sense that it is derived from the norms, and the purposes and objects that they stipulate that the agent's *actions* count as being directed towards. The intentionality of people's intentional states is derived: derived from the norms about what objects the person's actions count as being directed towards. The intentionality of the internal mechanisms, is similarly derived. Not by being implementations of the intentional states, but by being mechanisms that cause and enable the agent's actions to take place, and the objects those actions count as being directed towards.

Let's imagine that there is a particular neurological mechanism that enables me to write this particular paragraph, the one you are beginning to read right now. I've been intending to express the idea that this paragraph explains, for months, perhaps years. My desire to express it has led me to explain it (or at least its less refined "ancestors") to many people over the past while, and now I am writing it down. The internal mechanism(s) that enable me to do this, that enables me to talk about, refine and eventually write down this very idea, enables me to perform what Andy Clark would be happy to call a "representation-hungry" task. Some people would call this neurological state or mechanism or whatever it is a representation; a representation of that idea. Its content, then, is derived from the content of the intentional state that people felicitously attribute to me as part of the reason I write what I do right now. But note, the intentional state is a state felicitously attributed to me–a whole person–not to the internal mechanism. People attribute this state to me because I perform this activity of writing. People attribute to me the intention to get this idea down in writing. They say: "Mason has this idea". The action of writing this down is an expression of that idea. My acts of explaining it to various people were an expression of that idea. The intentional state is attributed to me as part of the reason for my actions. My actions *express* this content, as surely as my action of making coffee express my desire for coffee. The internal mechanism that enables me to express the idea I express, however, is not *itself* a representation of that idea. It is a mechanism that, by enabling me to perform certain idea-directed actions, plays a role in the actions I perform that are framed within; the practices that license attributions of content to my actions and license attributions of intentional states to me. Just as my queen in a game of chess is not *itself* directed at your king when I put you in check, but enables me to perform king-directed actions within the practice of playing chess, this internal state enables me to perform idea-directed actions, but is not itself directed at that idea. The neurological mechanism is not even an implementation of that content. It simply enables me to perform these actions, and by virtue of the content of the intentional state attributed to me (a whole person) because of those actions we can attribute a derived content to this mechanism as well.

Let's take a more practical example. I can plan my route for a excursion downtown using either a map printed on paper, or some sort of "cognitive map" implemented in my neurological structures. My *activity of planning* is directed at my route downtown, but neither the cognitive map or the printed map are *themselves*, independently, about anything. They only get intentionality attributed to them, by virtue of their *use*. They are used by me to perform route-directed activities. By virtue of their role in my activities, and in the practice of attributing goals and intentional states to me based on my activities and what they count as being directed towards ("I plan to take the 'high-level bridge' to get downtown" I say), the cognitive map and the piece of paper have derived intentionality. Independently of their role in such activities, neither the internal mechanism nor the printed map has any intentionality. It is only because they can be used as representations, that they have intentionality.

The same goes for expressions in our language. They have uses, and by virtue of these norm-governed uses, they play a role in people's world-directed speech acts. The speech acts count as being directed at various states of affairs, but the expressions themselves only have intentionality derivatively. They have "meanings" (better: they have *illocutionary act potentials*[113]) only by abstracting from all the particular uses that they can be put to, and generalizing about the kinds of things they enable speech acts to be directed towards. Without the phrase "New York City" in my lexicon, I couldn't perform speech acts directed at New York City. By abstraction and generalization, we can say that the phrase "New York City" is useful for referring to New York City. Its illocutionary act potential is such that it is good for performing illocutionary acts directed (in part) at New York City. This illocutionary act potential is held in place by the shared norms for using the term. Whether or not I have been to New York City, what I know about New York City, and even whether what I believe about New York City is appropriate to believe, these norms about the proper use for the expression make my speech acts using the word in a referring way, directed (at least in part) at New York City. [114]

Going beyond what kinds of speech acts an expression is useful for (or normatively should be used for), and talking about its meaning is a further abstraction and generalization. This may be useful for certain analytic purposes, for instances in talking about the elements of our symbolic system and how they relate to one another, but in order to explain how people are able to perform

[113] This phrase is William P. Alston's (1964), p. 34

[114] Putnam (1975a) makes a similar point. He argues that my statement "that's a beech tree" is either true or false, not because of the content of my own individual concept and whether that content correctly applies to the tree, but because of my linguistic community's norms for the correct use of the word "beech". The criteria for distinguishing elms from beeches exist in the linguistic community (anchored by the existence of experts who can rule on the truth of such a statement). Some words, says Putnam, require cooperative activity to use; their extension is determined socially, not individually. Putnam, however, resorts to talk about natural kinds that such terms designate rigidly, once their reference is fixed in this way. On my view, there can be different norms for different communities, and each community has as much right to call the kinds that its terms are used to designate "natural" ones. Likewise, there is room for debate about what a speaker moved from one community to another (e.g. me moved to Twin Earth) refers to when using such a community-relative term. (Do I refer to water or twin water when I ask for a drink of water shortly after my arrival on Twin Earth?) As Dennett (1995a, p. 408 ff.) suggests, there could well be an indeterminate time when it isn't clear whose standards should be used.

actions directed at the things they count as being directed at, we might as well stick to talking about what kinds of speech acts it is useful for.

Thus the intentionality that we attribute to neurological states and the intentionality that we attribute to linguistic expressions are both abstracted and derived from their uses. Their intentionality is *abstracted* from the kinds of actions they enable the agent possessing them to perform, and it's *derived* from the norm-governed practices in which such actions have their life. The basic type of intentionality, however, is that of people's actions (both linguistic and non-linguistic). It is instituted in the practice of giving reasons for actions: The intentionality of each action is due to the objects and states of affairs that people's actions count as being directed towards, as instituted in the intentional states that the practice of attributing reasons licenses us to attribute to the agent to explain their reason for performing that action.

In this chapter I've completed my argument that all intentionality is based in shared norms for appropriate ways to behave and appropriate ways to think and reason. In the epigraph at the start of this Chapter, Humpty Dumpty claims that words can mean whatever he chooses them to mean. "the question is," he says, "which is to be master—that is all." If I am correct, then neither Humpty Dumpty nor the words he uses is master. Both are constrained by and should conform to the norms instituted within the practice of interacting with others. At least they should conform, if they think of themselves as "rational", which they also should do. Humpty Dumpty is definitely a nonconformist at heart. But he still conforms to some minimal degree; for instance, when he uses "There's glory for you" to mean "there's a nice knockdown argument for you", he recognizes his obligation to explain to Alice what he's trying to make the expression he just used mean. The fact that Alice didn't understand, because he wasn't following the norms, means that he *failed* to make it mean that. Now *there's* glory for you, Humpty.

All that remains to be done to present a naturalized account of intentionality, then, is to present an account of how practices and the norms that are instituted within them arise, to get the whole thing off the ground.

Of course, the difficult step in this story is to present an account of how linguistic practices, and the practice of attributing intentional states to others arose. After all, without the ability to attribute intentional states to others, we would not be able to use language. And without language we would not be able to express the content of the intentional states we attribute to others. And without these practices and their norms, instituting the intentionality attributed to creatures that cannot themselves attribute intentional states, there would be no intentionality at all. It's to such an account that I turn in the following, final, chapter.

Chapter Six

The Evolution of Intentionality

The origin and the primitive form of the language game is a reaction;
only from this can more complicated forms develop.
Language —I want to say— is a refinement, 'in the beginning was the deed'

—Ludwig Wittgenstein [115]

6.1 *"Are we nearly there yet, Dad?" A sketch of where we've come so far, and a little white lie about how there isn't far to go.*

So far I have been defending the thesis that intentionality arises only because of norms and the practices the norms are instituted within. In Chapter Two, I argued that intentionality is a normative concept. A representation or intentional state having content, or a words' having meaning' is based on norms for what they should represent or mean (what kinds of things they should be used to refer to). Since Chapter Two I've been asserting that the way to give a "naturalized" explanation of intentionality, is not to reduce the intentionality of people's internal representations to non-intentional properties of physiological states, but rather to embrace this normative basis for intentionality, and to explain how the normative practice of attributing intentionality evolved. In Chapter Three, I argued that human contentful intentional states, language, and explicit norm-following behaviour arise out of tacit non-intentional norm-abiding behaviour (e.g. resulting from behavioural conditioning). Chapter Four gave a normative account of language and the both tacit and explicit norms and practices that structure human forms of life. Chapter Five showed how the

[115] Wittgenstein (1980), 31, a note dated 1937. (This is, of course, a reference to Johann Wolfgang von Goethe's *Faust*, line 1237.) I presented an earlier version of this chapter (along with bits of Chapter Four)as a paper, entitled "In the Beginning was the Deed", a t a conference on Naturalism Evolution and Intentionality, at the University of Western Ontario. There Christopher Olsen pointed out to me that Peter Winch also wrote a paper titled after this quote in Wittgenstein ("Im Anfang war die Tat" Winch 1981), coincidentally also presented at a conference at the University of Western Ontario. There Winch notes that 'In the beginning was the deed' is a quote that reverberates through many corners of Wittgenstein's philosophy (p. 176). This chapter, indeed this dissertation, picks up a different set of reverberations than does Winch's paper.

justification for attributing intentional states to people is due to the normative practice of giving reasons for actions. Within this practice, we are licensed to attribute intentional states to people as reasons for their actions, and we are licensed to expect people to be committed to acting in certain ways, based on the intentional states that are appropriately attributed to them. I also explained how whatever intentionality that attaches to internal neurological states is abstracted and derived from the role these states have in producing behaviour, and the practices in which such behaviour is situated. Their intentionality, like the intentionality that we attribute to linguistic expressions, is abstracted and derived from their uses; it's abstracted from the kinds of actions (including linguistic ones) they enable the agent possessing them to perform, and derived from the norm-governed practices in which such actions have their life.

So, now I've embraced a norm-based account of intentionality, it's time to give an explanation of how normativity and intentionality could evolve. This final chapter aims to show that this kind of explanation can be given, and to present an initial attempt at giving it. The first section summarizes research showing the way in which tacit normativity can arise. The latter half will show how human languages and the norm-following practice of attributing contentful intentional states as reasons for actions evolved out of such tacit norm-abiding practices.

Along the way, I'll give a critique of accounts of the evolution of human languages that miss the crucial role of attributing intentionality to others in linguistic interactions. I'll show how missing this aspect of language entails missing an account of a crucial step or two in the progression from simple biological machines and their purely causal interactions, to human beings and all their capacities. Even for those accounts that acknowledge the importance of mindreading abilities for human interactions, however, explaining the emergence of these abilities is often seen as a rather mysterious step, or as something that was simply "stumbled upon" or "figured out" without explaining how this might have occurred. The aim of this chapter is also to fill in some of the blanks here.

6.2 *Normativity and evolution*

Somewhere along the line, back in the prehistoric past many species have hit upon and perpetuated what Dennett (1995a, p. 78) calls a *Good Trick*, evolutionary speaking. By clubbing together in groups, they have a better chance of surviving and reproducing. The reasons this grouping works are multifarious: a lesser chance of succumbing to predation by sharing the task of looking out for predators (watching one another's back), by sharing the task of looking for cultivating and/or acquiring sources of nutrition, sharing the task of raising the young, and so on. Strategies that individuals were innately disposed to adopt were selected for, and individuals that had the disposition flourished.

Some of these strategies weren't simply selected for randomly, and genetically "hard-wired", but learned. Behavioural plasticity, the ability to have some of one's dispositions "re-wired" through feedback from the environment as one develops, is itself a rather ingenious Good Trick. It allows individuals to respond adaptively to changing circumstances. It enables a flexibility that can out-maneuver rigid genetic hard-wiring of one's behavioural responses. Through behavioural conditioning by feedback from the environment, certain (successful) strategies were reinforced and other (not so successful) strategies were not, or were negatively reinforced. We see even the simplest creatures nowadays being influenced by such mechanisms.

In evolutionary theory, a startling puzzle emerges with the emergence of such flexible adaptation to one's circumstances: how do we explain the fact that often it is the case that learned strategies are not the sole discovery of an individual, which die out with that individual. *Somehow* these learned strategies are passed on to the young, so that the next generation can benefit from some of the strategies for survival that the previous generations discovered. The nature of this "somehow" needs explaining.

A thesis attributed to Jean Baptiste Lamarck[116] argued that traits acquired during an individual's lifetime were somehow embedded in genetic code and passed on to successive generations; a theory that has since been discredited. The Baldwin effect has been variously reviled and applauded, but is now generally accepted as the solution to this problem. James Mark Baldwin (e.g. 1895, 1896, 1902) proposed a scientifically respectable alternative to the less respectable Lamarckian theory: that individuals pass on via their genes, not the acquired characteristics themselves, but rather their *ability to acquire* certain characteristics. Individuals of a certain species flourish when they learn to

[116] In none of the papers and books do I find any reference to actual papers written by Lamarck. People attribute this theory to him, however. For a discussion of the debate over Lamarckian theories, see Dennett (1995a, pp. 320 ff.)

perform a certain task, and so there is selection for the ability to learn that task (e.g. an oyster-catcher's ability to learn to open oysters; apparently infants learn by following Mom around and watching her as she does it). Thus the ability to learn that particular tactic is an advantage to individuals, and there is selective pressure—created by the social environment, not just the physical one— that selects in favour of individuals more innately disposed to learn that particular tactic.

An extension of this effect, where the group's social environment affects the selective forces at play, alters not the selection of individuals within a group, but selection of groups themselves. Particular herds of cattle, horses and antelopes, etc., flocks of birds, prides of lions, troops of monkeys and apes, adopt different practices. Having that practice, learning to behave in that way and to socially condition others to do the same, can take advantage of conditions that favour a particular group over a group that doesn't have such practices. Imagine two groups of genetically relatively identical individuals, one of whose members practice the "altruistic"[117] strategy of crying out warnings about predators to their fellow group-members, while the other group adopts an "everyone look out for themselves" strategy. (This is an "altruistic" strategy, because calling out warnings can be deleterious to the caller's chances of reproduction while benefiting the group. The individual calling out the warning can possibly call attention to itself, and be the likely target of the predator.) If members of a group assist one another, by watching one another's backs, the members of that group, and thus the group itself, can have an increased chance of survival, compared with a genetically similar group which adopts an everyone for themselves strategy.

Such practices are group-level phenomena. They require the participation of all, or at least a significant proportion, of the group. The problem is one similar in form to an iterated prisoner's dilemma. Individuals' strategies on when and whether to cooperate can benefit themselves at the expense of others, or can benefit the group, at the risk of possible disadvantage to oneself. If just a few members practice this strategy, and doing so is to their disadvantage, then these individuals (and thus their practice) could have a selective disadvantage within the group compared to their fellows who, when they notice a predator, hide in an appropriate way and wait for some "gullible fool" to attract the

[117] Altruism has been one of the disputed aspects of behaviour in the debate for group selection. See Sober (1984) for discussion.

predator's attention. The problem, then, is that such strategies only benefit the group, if most members of the group cooperate in the practice.

In the standard prisoners' dilemma cases, it seems that the larger the group of participants, the less likely that cooperative strategies will develop. When *pairs* of individuals interact repeatedly, cooperative strategies are likely to develop via reciprocity. Boyd and Richerson (1992) point to an extensive literature showing that in an evolutionary setting with a pair of participants, the equilibrium strategy tends towards the form, "cooperate on the first iteration, then cooperate only if the other also cooperates." In larger groups, however, the models that Boyd and Richerson present and refer to suggest that "the conditions under which reciprocity can develop become extremely restrictive as group size increases above a handful of individuals" (p. 173). Defectors, they argue (p. 174), even if there are only a few, increase the probability that others will defect. The strategy of defection, introduced into a population of cooperators, can easily spread if there are only a few defectors. In a group of non-cooperators, the chances of a cooperating strategy spreading beyond a cooperating pair, is small. As the size of the cooperating group increases, the chances of defection increase. Thus in large-group prisoners dilemma situations, the likelihood of stable cooperative strategies arising is small. Such strategies are rarely stable. Yet such cooperative strategies do exist, and appear to be very stable.

The solution to the problem is that in standard evolutionary situations, where individuals live together in a group, there are more options open than whether to cooperate or withhold cooperation. Reciprocators can retaliate against noncooperators, by doing more than withholding future cooperation. Many other forms of retaliation are open to conspecifics, from execution or banishment, to beatings, or to making them the subject of gossip. Boyd and Richerson's model shows that by *enforcing* cooperative behaviour, in punishing those who fail to cooperate, very stable patterns of cooperation can result in larger groups.

Haugeland (1982, p. 16, 1990, p. 404) argues[118] that by these mechanisms of conformism —that is, imitativeness (the tendency to copy others' behaviour) and censoriousness ("a positive tendency to see that one's neighbours do likewise, and to suppress variation")— the clusters that group together aren't

[118] The 1990 paper includes an almost verbatim reprint of this part of the 1982 paper. Citations and quotations are from both, unless otherwise indicated.

herds, but *norms*. As groups coalesce, the peer pressure will act as a kind of "mutual attraction" between the group's behavioural dispositions, he says (p. 404, p. 16). What the norms will be, how strict they are, and how many of them there will be is due to many factors, including chance. The only things that conformism ensures is that there will be groupings of creatures, and that within such groupings there will be norms: "distinct, enduring clusters of dispositions in behavioural feasibility space". These are not explicit norms, such that the creatures are aware that they are following norms, or aware of what the norms are. They do not follow the norms *because* that's what the norms are (this is a trait, recall, that Heidegger says only humans have; it requires a language in which to express the content of the norm). These are simply dispositions to behave in certain ways, brought about through behavioural conditioning according to the practice of enforcing norm-abiding behaviour.

The Baldwin effect shows how, in a situation where the norms don't change significantly over many generations, there can be selection for the "natural" inclination to abide by these norms. Even if the practices change, though, there can be selection for an innate predisposition to be good at learning the particular kinds of practices the group participates in, even if these change over time. For instance, the selection for human beings that are good at learning to participate in particular kinds of linguistic practices, combined with selection at the group-level for languages that are easy for youngsters to learn, would be the process that has resulted in the refinement of human languages to a small range of the possible space of languages, and in the allegedly "innate" grammatical faculties Chomsky (e.g. 1986) points towards.

This is also the reason that a relatively innate disposition to acquire the ability to attribute intentional states to others seems to be "hard-wired" into human beings. There is a significant body of research showing how either mindreading abilities or their development at the appropriate rate, seems to be rather innately "programmed". I'll talk much more about this literature presently, arguing that selection for individuals better able to participate in the kinds of interactive practices that involve attributing intentional states to others, has selected within human populations for individuals better able to learn and apply these skills.

These processes can enable a group, to very quickly speed up the process of adaptation to the environment, and the group can win out in the struggle for space in the evolutionary landscape. Thus even individuals with an excellent individual chance of survival, could have their genetic line eliminated, because the group of which they are members loses out. [119] (Even the most superior soldiers sometimes die in battles, when their side's tactics are inferior.) Similarly, genetically inferior individuals have a better than average chance of passing on their genes, if someone is watching their back for them, and the group they are supported by wins out.

Another advantage to groups of this mechanism of reproducing specific advantageous behaviours is speed of adaptation; a culture's practices, as Sober (1991) shows, can adapt at a much faster pace than evolutionary changes would produce. (Witness the drastic changes in human cultures over the last few generations.)

Although the thesis of group selection has been generally dismissed in the past, it has recently regained a large measure of scientific respectability. As David Sloan Wilson (1997a) represents the history of the group selection debate, the processes of natural selection were once assumed to operate on many levels. George C. Williams' (1966) critique, arguing that natural selection never operates an anything above the level of the individual, however, roused a consensus of dismissal in the 1960s that some take to have killed the debate. When talking purely in individualist terms, and seeing culture as simply an aggregation of individual interactions, group selection looks remarkable like selection for more advantaged individuals, writ large. Most group-level phenomena were redescribed in terms of aggregates of individual level phenomena. Altruistic behaviour, for instance, was a cornerstone of group selection arguments at the time, particularly Wynne-Edwards (1962) . According to Elliot Sober (1984, pp. 188-9), Williams' main argument against Wynne-Edwards, was that altruistic behaviour could in fact favour the individual making the warning cry, by for example, creating a flurry of activity, in which the signaller is the first to safety. Or perhaps the warning cries benefit the altruists offspring and close kin, thus being a form of parental care. Or perhaps the cry doesn't in fact expose the signaller to risk, because predators cannot localize the source of the cry. Such individualist arguments and dismissals of group selection still occur. Even recently, Wilson (1997, p. 2) reports, many textbooks and

[119] See Sober (1984) for a characteristically thorough discussion of the likelihood of such group selection mechanisms. For a relatively complete bibliography and debate of the question of group selection mechanisms, see Wilson and Sober (1994).

articles still refer to this consensus as a way of dismissing the thesis as worthy of consideration.

Through the 1970s a theoretical framework emerged that could stand up to these criticisms. After also standing up to empirical testing in the field, the idea that The forces of natural selection can operate at many levels (genes, memes, individuals, family units, tribes, species). The supplemental issue of *The American Naturalist* journal that Wilson's article introduces is devoted to exploring and applying (not to debating) the thesis that selection at many levels, including the group level, is a respectable scientific hypothesis with good empirical and theoretical support.

At the group level, the forces of natural selection can operate very similarly to the way they work at an individual level. There are two necessary ingredients for the process of natural selection, argues Elliot Sober (1991, p. 478): differential fitness, and heritability. In the individual selection model of natural selection, we have variations in fitness of individuals due to differences in the individuals' genotypes. The fitter individuals have a better chance of reproductive success, and pass their genes onto their offspring. In the cultural selection model of natural selection we see a similar process. Between groups, their different practices will give some groups a relative advantage over others, and the selective forces will favour the continued survival of the group whose practices better contribute to the survival of the group, and the further propagation of those practices. One generation's selective fitness is passed on to their students. In each case there is also a drift of traits or practices, caused either by genetic mutations or by alterations in the way the practice is taught or adaptations by the learner.

This last point shows, though, that the above model of cultural transmission is an under-representation of the advantages of the process, however. One reason is that practices can be modified by design. Individual learners and practitioners can make what they take to be improvements in the ideas they are taught before they pass them on to their students. Thus (depending on the strength of the conformism to the original practice and in the general suppression of variation) there can be what Boyd and Richerson (1985, p. 9) call *guided variation* as well as random variation in the nature of the group's practices. A further difference, as Sober (1991) points out, is that cultural transmission is not just vertical, but horizontal and oblique: traits are not just learned from one's parents, but from other members of the same generation and the other generation. Furthermore, as Sober explains, individuals can be exposed to a wide range of ideas and can pick and choose the more attractive ideas to adopt. The practices that are perceived to be better will attract more students. Thus the spread of ideas has a lot in common with a contagion.[120] (Sober says that this model of cultural change is tied both to the genetic theory of natural selection and to epidemiology.) Boyd and Richerson (1985) give a good sketch of the ways the forces of natural selection on individuals have analogs at the group level, and of the differences between the two processes.

A requirement —one I haven't yet considered— of societies where participation in normative practices is enforced, shows just how arbitrary a stable set of norms can be. Boyd and Richerson (1992) point out that if punishing is costly to the punisher (e.g. engaging in physical combat) then punishing itself is an altruistic act; something the individual does that is to the group's benefit but at least at the risk of being detrimental to the punisher. Natural selection should, therefore, favour individuals who cooperate in the practice, but who do not punish others. So why do individuals punish others?

The answer is that the practice of punishing is itself a normative practice that can be enforced. Boyd and Richerson argue that what they call a "moralistic" strategy —where in addition to cooperating and punishing non-cooperators, individuals also punish those who fail to punish— is a very stable strategy for a population to adopt. Their model institutes the practice of punishing those "not in good standing": those who have behaved according to the norm (cooperate and punish those not in good standing) since the last time they were punished, or since the beginning of their interactions.

Such a moralistic strategy, they argue, can cause *any* individually costly behaviour to become evolutionarily stable, even if it confers no advantage to the group. This means that the practice of conforming to *any* set of norms can arise and become instituted in a normative practice, if the society also sanctions individuals who should punish, but refrain from doing so. These norms of behaviour can be of no benefit to the individual, nor to the group, but will persist because of the practice of punishing those "not in good standing". This further supports my contention back in Chapter Two, that an evolutionary explanation

[120] Dan Sperber (1996) uses a similar epidemiology model to explain cultural transmission. However, as Andrew Sneddon (forthcoming) points out, by focussing on culture as the production and spread of representations, rather than of practices, Sperber's model suffers from an "unduly limited and inert picture of what culture is".

can *explain* how normative practices arise, without needing to *justify* any particular norm. Many of the norms can be completely unjustifiable, yet good evolutionary explanations for how they arise and are perpetuated can be formulated.

It seems then, that appeals to norms and their enforcement, as a foundation for a phenomenon, is to appeal to a foundation with a respectable scientific pedigree capable of establishing a "naturalizable" foundation for that phenomenon. The slow random "generate and see if it works" strategy implemented in the mechanisms of natural selection, brings about non-conscious norm-abiding behaviour. Norms of behaviour and their enforcement arise over time in groups of genetically similar creatures that group together. They arise through methods of selection within groups for individuals that practice conformism and censoriousness, and selection within an ecosystem for groups whose practices confer some relative advantage to the group over groups of the same species that do not abide by the norms of that practice. Norms arise through the process of natural selection, and norm-constrained behaviour is a real phenomenon, to which a naturalistic theory can appeal to.

6.3 *The need for an account of the evolution of intentionality.*

So tacit norms, abided by due to peer-pressure and behavioural conditioning can arise in populations, due simply to the forces of natural selection. I have argued in Chapter Three that this is a foundation for the ability to *consult* norms and follow them explicitly, and in Chapter Five that it is also the foundation for the practice of attributing intentional states to others (and to oneself). How these are built upon this foundation is still in need of explanation. I'm going to give the beginning steps in such an explanation in the rest of this chapter. I'll argue that a limited ability to attribute intentional states to others (awareness, desires, intentions), arose in creatures with a complex political structure and a large amount of prefrontal cortex development that was selected for, in selecting for socially astute individuals and the capacity for some level of abstract thought required to do well in a complex social and political community. This, I'll argue was the precondition for the practice of employing a system of intentional states and the relations between them, which itself was a precondition for the transition from the kind of indexical signalling systems that our ancestors once employed and all animals still employ, to symbolic human languages. Furthermore, I'll argue that the combination of these phenomena is what "bootstrapped" human beings' cognitive capacities such that we are now able to do the remarkable things we can do. (And remember, as I explained in Chapter One, that it's these *capacities to do things* that we are trying to explain in cognitive science.)

It is curious to me that often in the literature on the evolution of language, the practice of attributing intentionality hardly gets a mention. Neither do the more individualistic conceptions of the practice, such as the ability to *mindread* or, synonymously, the ability to adopt the intentional stance. Nor does the even more "Cartesian" sounding concept of having a *Theory of Mind* (ToM), as developmental psychologists refer to it, merit much mention in the literature on the evolution of language. There are several good treatments of the evolution of mindreading abilities, particularly in the work of Andrew Whiten (e.g. 1993, 1996b, 1998); however, Such treatments for the most part shy away from the link between mindreading and language. Of those theorists on the evolution of communication who do acknowledge the importance of the ability to mindread there is often little mention of how this ability evolved. Marc Hauser's (1996) *The Evolution of Communication* is a prominent example. Hauser acknowledges the role of ToM in children's development of linguistic skills (p. 594 ff.). He also notes the role of ToM in people's social interactions and language, particularly stressing the need "to understand *why* individuals utter such signals, and what this says about their beliefs and desires" (p. 651). Yet Hauser is conspicuously silent on the question of how such abilities evolved. Daniel Dennett, with his concentration on the intentional stance, has books on evolution (1995a) and on the differences between human and animal cognition (1996). Dennett acknowledges the link between adopting the intentional stance and the cultural props, especially language, that have enabled human beings to do what we now can do. Because he explores this territory fairly thoroughly, especially in the latter sections of *Kinds of Minds* (1996), he comes up with many of the pieces of the puzzle. However, he seems to be missing a few (particularly the notion of the shared practices, as I argued last chapter), and he never seems to put them together. For instance, Dennett moves from talking about how the ability to attribute particular intentional states would confer a predictive advantage for a member of a complex social and political community (pp. 124 ff.), straight to talk about the advantages of using language to makes explicit characterizations of what one is doing and what one is thinking (p. 127). He also relates the distinction between attributing a reason to an agent, and it being a reason *for* that agent, to that agent having the ability to create and manipulate external symbols. However, Dennett then only talks about this in a developmental sense (pp. 148 ff.), about how as a child learns a language the child learns to make "self-commentary" about their own actions. Thus children, Dennett suggests, eventually comes to understand themselves in terms of the labels they apply to features of their own activities (p. 150). Dennett seems to shy away entirely from explanations of how these abilities might have evolved, and all his explanations are phrased in terms of individuals figuring things out, rather than members of a community following norms, and adapting those norms to enable toe practice to develop.

An account of the evolution of the ability to think about other community-members' intentional states and how it relates to the *practice* of giving reasons is needed. I'm going to give one, or at least a sketch of the way one might go about giving one and some of the elements of such an explanation. To do this I draw connections between developing this ability, and the ensuing practice of attributing intentional states to others, and thus to see how others would attribute intentional states to oneself, and to the further ensuing ability to participate in linguistic exchanges where the reason for making an utterance is to affect someone else's intentional states, to get them to appreciate your reason for doing what you did.

Along the way, I'm going to contrast my account with two generally well-respected treatments of the evolution of language (Deacon 1997, Pinker 1994), neither of which recognize the centrality of mindreading to the phenomenon they are trying to account for. They view language simply as the communication of information about the communicator's internal states. Human languages, to put it a little tritely, simply do this better and more explicitly. The evolution of human languages is the evolution of *better communication*.[121] (the account I gave in Chapter Four shows my disagreement with this position; I'll make the contrast explicit in what follows.) Deacon at least recognizes a connection between mindreading and human languages, but argues that the transition to human languages precedes the ability to mindread. One of the points I aim to prove along the way is that this cannot be so.

[121] Terrance Deacon (1997, p. 28), for instance, describes the evolution of human language as the evolution of a "more articulate, more precise, more flexible means of communicating."

6.4 *Language as Communication?*

It would be difficult to disagree with the claim that human beings use language to communicate with one another. However, I disagree with the claim that communicating is *all* we use language for. This claim is implicit (and often explicit) in many accounts of what language is, and how we are able to use it. This perspective is particularly visible in some explanations of the evolution of language. For instance, Stephen Pinker (1994), and Terrance Deacon (1997) both discuss the evolution of languages and human linguistic abilities in terms of the evolution of better communication. They talk principally about human communication systems, and how they are different from the communication systems used by our distant ancestors and by animals today. Pinker describes the ability to use language strictly in terms of communicating information. The ability to use language, to Pinker, is "the ability to dispatch an infinite number of precisely structured thoughts from head to head by modulating exhaled breath" (p. 362).[122] Pinker and Bloom (1990, especially pp. 712-3) similarly argue that the primary selective advantage in possessing language is that it enables "the transfer of beliefs and desires" (p. 777). Pinker and Bloom further argue that "...communication of knowledge and internal states is useful to creatures [like humans' ancestors] who have a lot to say and are on speaking terms" (p. 712). Terrance Deacon's (1997) account is different from Pinker's in many important and insightful respects. However, Deacon also sees the communication of information about the communicator's internal states, predispositions, beliefs, social status, intentions, and so on as the purpo;e of both human language-use and animal signalling behaviour.[123]

As should be apparent from the account of language I gave in section 4.3, using language to communicate the speaker's knowledge and internal states, to transfer thoughts, is not the paradigm example of language-use, but a *special case*. It's a special case that does need to be explained, but taking it to be the paradigm case, in relation to which all other language use is to be explained (as communication plus some other stuff), engenders what I take to be a distorted

[122] See Pinker (1994) Chapter 11, "The Big Bang," for a full explanation of Pinker's views on the nature of language.

[123] Deacon expressed the function of animal and human "communication" in these terms to me in personal communication. Deacon discusses the development of human language in terms of the differences between "animal communication systems" and human communication (pp. 30-34 and p. 57-59), and the transition from "nonlanguage to language communication" (p. 340).

picture of what language is and what we use it for. To borrow a metaphor from Terrance Deacon,[124] it distorts the picture just as a description which takes porcupine quills as the paradigm example of animal fur, in relation to which all other animal fur is to be described, would engender a distorted picture of animal fur. Porcupine quills are a specialization that evolved from animal fur. To describe all other animal fur as "porcupine quills, plus or minus something," would be a limiting and distorting way to describe other animals' fur. In a precisely similar manner, communicating information about internal states is a specialization of a more general phenomenon: doing something in a way that makes your reasons for doing it recognizable to another. My thesis, then, is that explanations of the evolution of language need to concentrate on the more general phenomenon of interaction, and to describe the communication of information as a specialized form of it, rather than describing the general phenomenon in relation to this specialized form.

I believe that ignoring both the interactive nature of language-use and the dependence of language-use on mindreading, and focusing instead on communication, is the reason that many theorists see such an insurpassably vast difference between human languages and the sort of signalling systems from which human languages must surely have evolved. The difference between human languages and non-human signalling systems is seen by some, most notably Noam Chomsky and Stephen Jay Gould, to be such a vast difference that they have gone so far as to deny that an evolutionary story can be told about how we humans came to have such sophisticated languages.[125]

I disagree. I think we can give a coherent account of how human languages evolved out of more primitive signalling systems, but only if we view the use of words and signals as a refined and sophisticated form of purposeful interaction. When we take these aspects of language-use into account, it's easier to see how human languages are similar to, and could have evolved from, the sort of simple signalling systems animals use, while at the same time also being

[124] Deacon used this metaphor (in personal conversation) to make a quite different point, and I do not mean to imply that he would endorse the use I make of it here. Deacon has informed me (personal communication), in response to a rough description of my use of it, that he probably wouldn't do so.

[125] See, for example, Chomsky (1988), p. 167 ff. and Gould (1989), p. 14. Chomsky and Gould both maintain that human language capacities must have evolved as a side-effect of selection for other capacities, perhaps for increased brain-size. See also Pinker (1994, ch. 11 "The Big Bang") and Dennett (1995a, p. 384-400), for more thorough discussions and criticism of Chomsky's and Gould's positions.

able to see how different our linguistic systems have evolved to become. On this approach, as I'll soon show, the difference between human linguistic interactions and the sort of animal signalling interactions we observe in the wild, is a very large difference in sophistication, but not a sharp discontinuity in kind. Before I discuss the details of this account, I want to look at a couple of accounts that disagree with Chomsky and Gould for different reasons than mine.

In order to explain how we evolved from the kind of primitive signal-users that most animals are now and that our ancestors undoubtedly once were, to become the language savants humans are now, we need a platform from which both the similarities and the differences between animal signalling and human language are visible. Viewing both animal signal-use and human language-use in terms of interaction, rather than communication, provides such a platform. The important similarity, as I'll argue, is that both animal signals and human speech acts are actions performed in order to induce another to respond in a particular way. (Of course, in most animals this "in order to" is quite tacit.) Natural selection in favour of the ability to make a noise or a gesture that induces another to act to your (sometimes mutual) advantage, both perpetuates animal signalling behaviour, and is also the force behind the evolution of human languages.

The interesting difference between animal signalling and human language-use is that animals interpret the signal itself, while (as I argued in section 4.3) humans can and do look to the reasons why a speaker uttered what they did. To properly understand a person's utterance, as I argued, the interpreter must interpret not just the words uttered, but the speaker's *reason for* uttering those words in this situation: Why did that person use those particular words? What did they intend to do by uttering those words? What normative practice did they intend that action to be situated within? What did they intend me to do in response? I'm going to use this platform as a beginning point, to show how such human linguistic interactions could have evolved out of such signalling interactions.

6.5 Language and Mindreading

Back in Section 4.5 I talked a lot about autism, and how autistic people's deficits in mindreading abilities can account for some of the peculiarities of the linguistic and social interactions. Terrance Deacon (1997) disagrees. He concedes that autistic people do have "a quite specific difficulty with imagining what is going on in other people's minds" (p. 275), and also have difficulties with taking other's perspectives and appreciating that others may possess different information than themselves because of their different perspectives. These deficits, he says, would be predictable with people who have damage to their prefrontal cortex, as autistic people appear to have (p. 275). However, Deacon dismisses an impairment of theory of mind or of social cognition as an explanation of the various deficits and behaviours demonstrated by autistic people. Deacon argues that such impairments "cannot alone explain the failure to develop normal language..." (p. 274).

This contention appears to be the result of Deacon's conception of what a "normal language" is, and just what abilities support the development and exercise of normal language. Deacon views learning a language as learning to use what he calls a "symbolic" system of communication. (By "symbolic" he makes a contrast with indexical and iconic, in the Peirceian sense I introduced in Section 2.4.) I contend, however, that an impairment in the ability to mindread can explain the deficits in autistic people's linguistic abilities (as well as in their social abilities). But this will only be apparent if we view language-use, not simply in terms of the communication of information as Deacon does, but in terms of a norm-governed interaction which depends upon recognizing others' reasons for acting, and making your reasons recognizable to others.

Deacon's position on the nature of language is used to support his theory on the evolution of language. In what remains of this chapter I'm going to the interactive approach to language to do the same. While Deacon highlights many important and interesting aspects of the nature of language, I believe that the embodied action approach to language-use that I have been arguing for contrasts with Deacon's in some important ways on the evolution of language.

Deacon and Pinker offer insightful accounts of human language and the cognitive mechanisms language-use depends upon, and offer suggestions as to how we human beings evolved to be such "language savants" in comparison to other animals. However, I believe that this focus on language and signalling as mediums used for the communication of information blinds them both to the ways that interaction and mindreading supported the evolution of human languages and the evolution of the linguistic abilities each puts at centre-stage.[126]

[126] Curiously, in the preface to his book, Deacon states as his principal research question: Why are there no simple non-human languages? Deacon argues that complexity of human

Deacon summarizes well the commonalities between himself and Pinker:

> Both Pinker and I argue that a very simple protolanguage could have evolved in an early hominid ancestor in the absence of any specific language adaptations of the brain, and that the adaptive advantages of language communication would have subsequently provided selection for progressively internalizing certain crucial features of language structure in order to make it more efficient and more easily acquired. (Deacon 1997, p. 328).

Thus both argue that language and brain structure co-evolved, brain structures being shaped by selection in favour of people able to learn and use languages (for communication) more easily, and languages getting more powerful and complicated, and easier to learn for people with brains like that.

Pinker and Deacon differ, however, in just what "crucial features" have been internalized in human brains through this co-evolution of language and brain-structure. Pinker argues for an innate "language instinct", one that is composed of many parts:

> syntax, with its discrete combinatorial system building phrase structures; morphology, a second combinatorial system building words; a capricious lexicon; a re-vamped vocal tract; phonological rules and structures; speech perception; parsing algorithms; learning algorithms. (Pinker 1994, p. 362)

All these parts, he says, are physically realized in "intricately structured neural circuits" the foundations of which have been "laid down by a cascade of precisely timed genetic events" (p. 362). Pinker argues that these innate faculties, or the genetic "programming" that leads to their inevitable development at the

language cannot be a barrier, since there are no simple animal languages either. By a simple language, Deacon asks us to imagine "a language that is logically complete in itself, but with a very limited vocabulary and syntax, perhaps sufficient for only a very narrow range of activities" (p. 40-1). Deacon imagines that the first human language could have been a simple system just like this. Co-incidentally, this is precisely the sort of "primitive" language Ludwig Wittgenstein (1958) depicts in the beginning few sections of the Philosophical Investigations. (a book which does not feature in Deacon's bibliography.) Wittgenstein's point there is a major influence on the perspective I articulate here: that we cannot analyze such languages, as Deacon does, in terms of "a mode of communication based on symbolic reference (the way words refer to things)" (Deacon 1997, p. 41); we must instead study languages, even primitive languages like the language of his tribe of builders, as inextricably embedded in a form of life. That is, we ought to study language-use as a mode of interaction in which linguistic tools are used in interactions with others as part of social practices.

appropriate time in a child's development, are now the genetic inheritance of every healthy child.

Deacon disagrees with Pinker, in that he maintains that many of these mechanisms couldn't have evolved due to natural selection for the abilities they enable. He argues that "The relative slowness of evolutionary genetic change compared to language change guarantees that only the most invariant and general features of language will persist long enough to contribute any significant consistent effect on brain evolution" (p. 329). Deacon argues that many of the features Pinker believes to be innate depend on features of our languages —their grammars— that haven't been invariant long enough to select for the mechanisms Pinker concentrates on —in particular the Chomskyan "universal grammar faculty". (Deacon argues that languages themselves have co-evolved with brains, to be easy to learn for the kinds of brains that human language-learners have.)

I'm not going to argue here about whether Deacon's criticism of Pinker is effective. The important point for present purposes is that Deacon criticizes Pinker for not looking at "invariant and general" enough features of our languages. Deacon concentrates on what he takes to be a more general and invariant feature of our languages: they are all "symbolic" communication systems, in contrast to the "indexical communication systems" non-humans use.

I introduced Peirce's distinction between symbols indices and icons back in section 2.4. Recall that for Peirce, the distinction between indices and symbols was that indices are interpreted to have a causal, existential connection with their objects, while symbols are interpreted according to conventions for interpreting that type of symbol. Deacon draws on Peirce's system, but with an interpretation of a symbolic system as having a "conceptual role semantics" flavor that I find rather questionable.[127] First Deacon's explanation, then why I

[127] In a review of Deacon's book, David F. Armstrong (1998) also points out that "semioticians may be troubled" by Deacon's interpretation of Peirce's system. However, Armstrong says that Deacon's "use of it to elucidate issues in behavioural evolution is so clearly explained that he should be forgiven any possible misinterpretation of Peirce's exceptionally murky theoretical pronouncements. The underlying ideas concerning the nature of symbols, indices, and icons are highly valuable and are put to good use by Deacon" (p162). While I agree that Deacon's use of it is very clearly explained, I disagree that the misinterpretation is harmless. It may be harmless with respect to Armstrong's critique (that Deacon doesn't take the gestural origins of language seriously enough). However, it is at the heart of my troubles with Peirce's account of the evolution of language.

find it a questionable interpretation of Peirce. The part that he misses of Peirce's system is a rather important part for my purposes.

Vervet monkeys' warning calls are a good example of an indexical communication system for Deacon; they apparently make one warning call for eagles, another for snakes, and another for leopards. Each of these warning calls is used to warn of a different predator, and induces the caller's fellow monkeys to engage in the type of avoidance behaviour appropriate to that predator. The calls are purely referential. There are no conceptual relationships between different calls; the only salient feature of the eagle warning call is that it warns of a particular eagle or eagles. Each time it is used, it refers to a specific instance of an eagle.

In contrast, a symbolic communication system (see Deacon (1997) pp. 79-92 for a more thorough explanation) is a system in which the words or "symbols" are conceptually linked together in a complex network. The word's meanings are linked together like the interrelationships between the entries in a dictionary (p. 82). In addition, the words are general types, that have referential links to types of objects (rather than particular tokens of that type, as is the case for indexical systems). The different types of referents are also related together in a complex network of physical relationships mirrored in the conceptual relationships between the words. The conceptual relationships between the meanings of the words have primacy in this system; the referents can even drop out altogether. Thus I can know what the terms "electron microscope", "Santa Claus", "Nepal", "maternal great great great grandmother", and "neutron star" mean through the conceptual links these terms have with other terms, in spite of the fact that I have never encountered the referent of any of these phrases, and the referent of some of them doesn't exist. The power of a symbolic system, says Deacon (p. 83) comes from its "*combinatorial* possibilities and impossibilities". This is why words need to be in combination with other words, he says, for them to have any determinate reference. He says that symbolic reference depends "on combinations both to discover it (during learning) and to make use of it (during communication)" (p. 83). This *combinatorial system,* is the distinguishing feature of symbolic human languages.

To Deacon, the evolutionary move from non-language to language "communication" requires un-learning a set of very useful indexical associations between words and objects in order to adopt an even more useful symbolic communication system (p. 92 ff.). Symbols cannot be learned one at a time, says

Deacon. The acquisition of symbols is the acquisition of a symbolic *system* (p. 92). Symbols can only be learned one at a time when they are being added to an already discovered symbolic system. Moving from using an indexical to a symbolic communication system would require some abstract conceptualization abilities he argues. One must be able to conceptualize the advantage of adopting the higher-level even more useful symbolic system and in suppressing the set of previously learned indexical associations. This requires a jump across an evolutionary "chasm". A species will not stumble across such a chasm. The "symbolic *insight*" (p. 93, Deacon's emphasis) must be leapt across intentionally; it must be "*discovered* or perceived, in some sense, by reflecting on what is already known" (p. 93). That this discovery requires a great mental effort and well developed abstract conceptualization abilities, he argues, explains why no animal but humans have been able to make the transition.

The major problem that I have with this interpretation, is Deacon's focus on a symbolic system as an abstract system of signs, simpliciter. The *users and interpreters* of the signs are almost irrelevant to Deacon's account of what makes a sign a symbol. When Deacon discusses the way a sign's interpretation (the "different patterns of mental action", and the "difference in the interpretive process" (p. 65)) makes it the type of sign it is, he does point out that it is a user's response that determines how the sign determines its reference. For Peirce, the difference is in whether the interpretation is made on the basis of a relation between the sign and object based on resemblance, on causation, or on convention. For symbols, the relation between the word and the interpretant is sustained by a convention. Back in Chapter Two, I interpreted this convention in terms of the *shared normative practice* of using the word to perform certain kinds of speech acts (the word's "illocutionary act potential"). Deacon, however ignores this talk of convention or norms, and goes straight to talking about connections between symbols themselves:

> The symbolic basis for word meaning is mediated, additionally, by the elicitation of other words (at various levels of awareness). Even if we do not consciously experience the elicitation of other words, evidence that they are activated comes from priming and interference effects that show up in word association tests. (p. 64)

The *conventional* underpinnings of the process of interpretation, that for Peirce underlies the nature of these associations between words and the mental acts that are their interpretants is entirely bypassed.

The basis of the relations between words that Deacon points to as the basis of the interpretation of symbols, is for Peirce, a shared set of norms about what concepts one should associate with a word, or (as I would prefer) in terms of a shared set of normative practices of using the word to perform certain kinds of speech acts. This is central, for Peirce, to the nature of a symbolic system. The basis in shared norms being missing from Deacon's account, however, is indicative of Deacon's conception of language, and as I will soon argue, of his mistaken conception of how language and mindreading evolved.

This conception of a language as using a symbolic system to mean something by one's utterances, for Deacon, is something that *one solitary person could figure out and employ*. As I mentioned above, for Deacon, a symbolic system must be figured out or discovered by someone with enough abstract conceptual abilities to appreciate the power of the system. As Deacon describes it, one solitary person could do this. Even if Deacon is correct, and one solitary person could discover the power of a symbolic system, it's the *shared* nature of the symbolic system that sustains the referential links between signs and their objects. And it's being a *normative* system is what ensures that the system is shared. Furthermore, the system being discovered by one solitary creature isn't much use as a system for communication (as Deacon conceives it to be), unless some other creature can interpret the signs in the same way. These points will be important in the next sections, when I talk about how such normative language systems and practices evolved.

Before I get to that, however, I should talk about what such systems evolved *from*. Both Pinker and Deacon agree, there is a sharp discontinuity between human communication and animal communication. This discontinuity is so sharp that the difference between language and non-language communication now appears to be a vast, almost unbridgeable chasm. Only human brains have evolved the language-specializations (they disagree about what these are) that make it so easy for human children to learn human languages. And this also explains why it's so difficult to teach even the most intelligent animals even a simple human-style symbolic language.[128]

[128] Deacon does make exceptions for chimpanzees, such as Sherman and Austin, who eventually learned to use a symbolic system, but only after extensive training (pp. 84-98). In fact, Deacon uses Sherman's and Austin's cases as an illustration of the difficulties of acquiring a symbolic communication system. (This interpretation of what Sherman and Austin learned is also contentious, but I do not wish to debate the matter here.)

There are many indications that mindreading abilities were a selective advantage to their possessors, such that once it was discovered, it was a significant factor in the (reproductive) success of those individuals who possessed these abilities and of groups whose members' practices utilized these abilities. One is that there are indications that autism is a genetic disorder. Rutter and Bailey (1993,) claim that "findings from both twin and family studies show that autism is *genetically* associated with both language and social abnormalities, and especially with their combination" (p. 487, my emphasis) (see also Rutter 1991). This hints strongly that there has been selection within the human population for the ability to mindread (or the ability to learn to mindread). Thus while mindreading abilities are part of every normal human being's genetic heritage, they can still fail to be passed on in certain rare cases.

Another indication that mindreading is an evolutionary adaptation is that we share this ability, in a far less refined form, only with our closest non-human relatives, chimpanzees. While summarizing research on chimpanzee mindreading, Andrew Whiten defends this focus on chimpanzees by arguing:

> The closeness of our relationship with chimpanzees is only one reason for focussing on them here, however; they are simply the only apes for which any significant evidence of attribution of mental states exists (Whiten 1993, p. 373).

Much primate mindreading research has shown that chimpanzees can act in ways that human interpreters take to be directed by utilizing mindreading abilities. They have been observed to behave differently to others, based on differences in what that other has seen, what others know, what they are likely to desire, and what others intend to do.

Chimpanzees have not yet been shown to pass anything like a false belief attribution test, however (1993, p379). But Whiten suggests that the work has hardly yet begun, and that this may be largely due to the fact that such a result is seriously impeded by the difficulty of designing such a test for non-linguistic creatures. Evidence from the use of tactical deception in chimpanzees is likewise negative, so far. Whiten and Byrne (1988) reported cases where chimpanzees deceive one another. This paper is often misinterpreted, Whiten claims, as evidence for the *intention* to create a false belief. However, no such conclusions are (yet) warranted, he argues (1993, p379). The behavior may have this effect, but the intention (and thus an awareness of false beliefs) has not been demonstrated, yet. Whiten argues (while citing Chandler, Fritz and Hala 1989) that this would require a chimpanzee, after one deceptive tactic fails, to try other alternatives, "the one common factor between them being their potential to create a specific false belief" (p. 279). No such behaviour has been recorded yet. I argued in Chapter Five, about attributing *contentful* intentional states (even to oneself) requiring a shared public language in which to express the *content* of intentional states attributed. Based on this conclusion, I am pessimistic that evidence of false beliefs in non-linguistic creatures will ever be found.

Chimpanzees, then, can be seen to use an *indexical* system of interpretation of others' behaviour. They interpret others behaviour by responding differentially to conspecifics (and human experimenters) based on what causes them to behave in these ways. As I argued in Section 2.5 symbols are the only kind of signs that have contents. Indices are based on interpreting cause and effect relationships between the object and the sign. Thus events happen in front of another, and that other is caused to behave in certain ways. The observer mentally labels the people who are likely to behave in those ways. People the observer has labeled in this way the observer can expect will respond in certain ways to particular events. For example, in chimpanzee those males who have recently groomed high-ranking females are often helped by those females when disputes break out. By keeping track of those males the observer knows have been recently grooming the high-ranking females, the observer knows who is likely to be helped in disputes, and thus who not to pick on. Keeping track of tacit "labels" that are causally (not contentfully) associated with a property such as *recently groomed a high-ranking female* is an important political and social tactic in chimpanzee societies. Similarly a label that is associated with a condition that we might describe in a more "intentional" way, such as *has seen me find this food* could be rather useful to apply and keep track of. This would be especially useful if one occasionally would like to keep tasty found morsels to oneself, but also knows that others could administer a beating to get the food if they know about it. If all the others to whom this condition applies are inferior on the hierarchy, then one is safe to hide and consume it alone. By using its limited abstract thought capabilities to keeping track on the things that others know, want, intend to do, and so on, a chimpanzee can have a distinct advantage in a the kinds of complex hierarchical political societies typical of chimpanzee troops.

That such "ground level"[129] mindreading abilities, in a primitive (contentless) form, were possessed by the pre-australopithecine ancestors we share with chimpanzees does require an assumption that mindreading in chimps and humans are homologous, rather than an analogous traits, however.[130] Analogous traits are ones that perform the same function, but evolved independently, in different ways; an example is bats' and birds' wings; bats' mammalian "forearm" bone structure shows that bats evolved wings independently, not through sharing a common winged ancestor with birds. The trait is simply useful enough that two species stumbled onto it independently. Mindreading abilities in humans and chimps would be analogous if each species evolved this ability independently, after our evolutionary paths separated, just because it's an incredibly useful ability. Homologous traits, on the other hand, are ones that developed from the same point, even if they may not now share the same function. The aforementioned "mammalian" forearm bone structure possessed by dolphin's flippers, bat's wings, cows' forelegs and human hands is a good example.

Stephen Pinker (1994, p. 349) points out that it would be interesting to find out whether "human language is homologous to anything in the modern animal kingdom". If we found such a homologous trait, it would be because a common ancestor of each species possessed the precursor of this trait. Pinker, with his focus on grammatical ability and communication as the important aspects of language-use doesn't think there is such a homologue of human language in the animal world. However, in addition to these grammatical and lexical abilities, the ability to mindread is also an important aspect of language-use, and is an ability that we do share (in a this contentless, indexical "ground-level" form) with chimpanzees, our closest non-human relative. Since we share so many other genetic characteristics, the idea that our common ancestors were able to mindread in this tacit, contentless way, and that after our evolutionary paths diverged, humans simply developed language and "top floor" mindreading abilities, on top of these ground-level mindreading abilities in a way that chimpanzees failed to does seem to be the better explanation.

[129] Andrew Whiten (1996b, p. 291) introduces this term to talk about "ground level mentalists" whose conception of mental states is rather inexplicit; they have the capacity to act with respect to others' mental states, but not necessarily to represent them as mental states.

[130] I've based this distinction between evolutionarily homologous and analogous traits on Stephen Pinker's (1994) explanation, pp. 347-48.

6.6 *Signals: for communication or manipulation?*

While Deacon criticizes Pinker for not looking at "general and invariant" enough characteristics of language, I in turn want to say the same of Deacon. Being able to use indexical or symbolic signs may be important, what we use these signs *for* is even more important, and even more invariant and general. We use these systems in combination with mindreading abilities to perform speech acts that aim at inducing others to respond to what we do.

Animal signals do something very similar: they do not communicate information to others, but manipulate others' behaviour. Manipulating others behaviour, getting them to respond in a particular way, I argued in Chapter Four, is the reason most speech acts are performed. It is also the (tacit, selected-for) reason that animals employ signals. As I have been explaining, Deacon, Pinker and many others maintain that both animal signalling and human language are used to communicate information to others about the signaller's internal states. However, as Dawkins and Krebs (1978) argue, animal signallers who simply communicate their beliefs and intentions to others would be open to manipulation by others, and would not gain any significant advantage for themselves. Communicating to others exactly what you want and intend to do would only be an advantage, perhaps, in a completely cooperative environment where there is no competition for scarce resources.

Abilities that give others an advantage over you, like the ability to communicate information about your wants and intentions, are not generally selected for. Unless your social group is completely cooperative, the ability to communicate your beliefs and desires and intentions to others confers no particular selective advantage, per se. The ability to alter others' behaviour in ways that benefit yourself, however, does confer a selective advantage.

The ability to use signs and signals as tools for manipulating others' behaviour to one's own advantage is the sort of ability which is selected for. Dawkins and Krebs argue that animal signal-use is more akin to advertising or propaganda than communicating information. Animal signals aren't used simply to inform others about the signaller's internal states, but to prompt or provoke others to behave in ways advantageous to the signaller.[131] For instance a bird singing in a tree isn't communicating the information that he has built a

[131] This ensuing behaviour brought about by an animals signal can be either competitive, mutually beneficial, or co-operative; see Krebs and Dawkins (1984), p. 391

nest and is ready to mate. His song is more the equivalent of an advertisement for his genes: he is advertising the fact that he can sing so loudly and yet not get eaten by the predators his song can attract, and calling attention to the marvelous nest he has made. The purpose of such a song is to attract a female to mate with him and not with someone with inferior abilities (and thus inferior genes).

Pinker and Bloom (1990, p. 777), however, reject this kind of analysis—in my view unfairly. They dismiss a peer commentary by Charles Catania (pp. 729-731) that language is used as part of an interaction, where the speaker's purpose is to effect some change in the hearers' behaviour. Catania points out that if communication has any selective effects, these are only because communicating can change what someone else does. Pinker and Bloom counter that while language

> "is related in some way to changes in others' behaviour, its proximal effect is the transfer of beliefs and desires, and any causal influence on behaviour is so circuitous, indirect and unreliable that it makes no sense to identify manipulation as a principle selective force in its design." (p. 777).

Thus Pinker and Bloom concentrate on the transfer of beliefs and desires as the proximal, and thus to them the most important, effect of someone's utterance. Pinker and Bloom's characterization of the effect of a signal as "the transfer of beliefs and desires" raises problems. I can ignore the problems with the ill-chosen use of the word "transfer" here; implying a copying of the signaller's desire in the signalee, as opposed to *inducing* certain (perhaps different) beliefs and desires. However, I also have a problem with the use of "beliefs and desires" here. This implies that the signals operate at a much more explicit, intentional level than is warranted. Most of the beliefs and desires induced in the signalee would be completely tacit. Most animal signallers and signalees simply behave in ways that are *describable by mindreading humans* in terms of their beliefs and desires. However, descriptions in terms of the way the signal induces certain behaviour in the signalee would seem to be more warranted.

My principal problem with this objection is similar in form to my objection back in Chapter Four, to Davidson's concentration on *first meaning*. Pinker and Bloom concentrate on the means, they treat the end it's a means to as of only incidental importance. But Catania is correct here. Altering someone else's beliefs cannot confer any selective advantage, unless those beliefs alter the others behaviour in favourable ways. Pinker and Bloom ignore the fact that the utterer' (often tacit) reason for uttering what they do is not simply to give others information, but to do something which will affect (ideally advantageously) the way those others behave. (Even in human language-users, this is done by getting others to attribute the hoped-for response that your action was a means of eliciting, as your reason for performing the action). Giving others information about your beliefs and desires, as opposed to manipulating others behaviour, is not the kind of ability that confers a selective advantage. In fact, it confers a selective disadvantage.

6.7 The evolution of "ground-level" mindreading

Pinker and Bloom also dismiss Catania's commentary by claiming that such an account "posits manipulation without explaining why nullifying countermeasures were not evolved by the manipulees..." (p. 777). Although Catania doesn't make this point, evolving such countermeasures against manipulation —or "sales resistance" as Krebs and Dawkins (1984, p. 391) put it— is an essential factor in the evolution of human mindreading and language.

Krebs and Dawkins (1984) argue that signals being employed to manipulate others' behaviour to the signaller's advantage creates conditions under which the ability to mindread would be very advantageous.[132] In situations where someone else uses a signal to influence your behaviour to their advantage, it would be a distinct advantage to be able to recognize the signaller's intentions —to be able to think about *why* the signaller might give you that signal — as indicated by contextual clues and by their facial expressions, by other non-signalling "body-language" (like blushing, for instance). For example,[133] if I give the signal normally used to induce someone to come closer (e.g. beckoning), while at the same time my facial expression and body language (combined with the context and history of interaction between us) betray

[132] I should point out, however, that Krebs and Dawkins use the term "mindread" in a different sense to the one I've been using. They use it to refer to what (Whiten 1996b) calls "sophisticated behaviour-reading". They use it as "a catch-word to describe what we are doing when we use statistical laws to discover what an animal is going to do next" (p. 386). Their description of the manipulative nature of signalling, and the defense that mindreading (in my sense) gives against manipulation, nonetheless still applies especially well here, but to a more limited range of creatures than the wider application they make of it.

[133] This example is adapted from Whiten (1996a).

aggressive intent, if you're able to recognize my aggressive intentions you'll be less likely to fall victim to the surprise attack I intend to mount.

Even very "ground-level" mindreading abilities enable some conniving deceptive tactics to be used, but also provide defense against such deceptive signalling. One often-cited anecdote[134] in primate deception literature gives an excellent example of the advantages that mindreading confers a social animal in a complex hierarchical, norm abiding (with punishments for transgressions and reciprocation of cooperation) social environment such as a chimpanzee society. In the example, experimenters hide food in special bins around the forests where a troop of chimpanzees live. As the story goes, an inferior chimp was about to look in the food bin to see if any food had appeared, but as he was about to open it, saw that the local "bully" was watching him. He noticed that there was food in the bin, but didn't remove the food, but pretended that the bin was empty by closing the lid again and walking away. He intended the bully to take his closing the lid as a sign (indexical) that there as no food there. When he thought that the bully was gone, he went back to the food bin to take the food, only to be assaulted by the bully who had hidden to see what the smaller chimp did next. The bully had seen through the inferior chimp's signal. He had recognized that the chimp had seen food in the box (facial expression might have been a giveaway—chimps have very expressive faces), and had thus recognized the young chimp shutting the lid without removing the food as a "signal" (indexical) performed to cause him go away, as he would if there was no food there. The young chimp handed the bully the food (usually a way of losing the prize while avoiding a beating). The bully took the food, and administered a beating anyway, presumably as retribution for the attempted deception (or perhaps better: for the attempt to make him go away, when he shouldn't go away).

It is definitely in the signallee's interest to be able to interpret such manipulative signals with an eye past the way one would standardly respond to the signal itself, to think about the signaller's intentions in signalling; to be able to recognize the purpose for which they signalled in that way, in this context, to me, with that social relationship between us. Furthermore, having beliefs about what the signaller is aware of and what the signaller wants would further enable

[134] I cannot recall where I saw this situation described (it may have been in a talk by Daniel Dennett). However, whether it is true or not isn't important. Even if it's not true, it is a good example of the kind of situation where mindreading abilities protect one against being manipulated.

the signallee to see past the signal itself to the signaller's intentions in so signalling.

Acquiring the ability even to mindread at this tacit level is not a trivial feat, however. As Deacon (1997, p. 427) notes, it involves attending to an abstract attribute, since someone's beliefs and intentions are not immediately visible to others. All that are immediately visible to others are their behaviour, and what's happening around them. Without the ability to mindread, all that would be available to an agent would be a very large number of cause and effect style associations between others' observed and predicted behaviour (predictions that may be relevant to one's own future actions).[135] For instance, such ground-level mindreaders –as I described earlier– can label certain individuals according to things that they are aware of (e.g. X saw me close the lid on the food bin) they can associate these labels with future likely actions (e.g. X will go away). But in complex social environments many other events might also cause X to go away from the food bin (e.g. X saw a more powerful chimp take all the food out). Additionally, any other actions are also predictable of others if this label is attached to them (e.g. X will beat me up if he recognizes that I'm trying to deceive him). In such situations, it would be very advantageous to find a way of simplifying all the cause and effect associations relevant to the situation. For example, imagine that eight different types of event involving individual X were each predictive of a certain range of seven behaviours expectable of X in response to various situations. If this were the case, then if one wanted to be able to predict how X is likely to respond to those situations, one would have to remember 56 different cause and effect associations between types of event and types of likely behavioural responses. However, positing an abstract intervening label between these, such that the eight types of events happening around X were associated with that abstract label (which gets attached to X), and if attaching that label to someone was predictive of the seven different responses that person is likely to make, then only 15 associations would have to be memorized. Doing so, then, can effect a considerable savings in the associations to remember, and also in the number of different labels one has to keep track of. Thus proposing abstract intervening labels that (for us language users could be interpreted as *X knows that there is no food in the box*, or *X wants meat*), can tacitly

[135] Andrew Whiten (Whiten 1993, p. 385 ff., 1996b, p. 283 ff.) gives good examples of the cognitive economy gained in positing intentional states as abstract intervening variables between the observed and expected behaviours of others.

serve to guide behaviour around X, while being considerably less cognitively taxing. The ability to attribute simple intentional states like this to particular individuals would confer a significant advantage; both in the number of individuals one could keep track of, and the number of these more useful labels one could learn to associate with various behaviours and predicted responses.

6.8 *The co-evolution of "top floor" mindreading with language, brain structure, and norms.*

Deacon (1997, p. 422) points out that positing such an abstract intervening variable probably could only be done by a creature which possessed a relatively developed prefrontal cortex (which enables sufficient capacities for abstract thought). Deacon argues, however, that the ability to represent another's mental state, as opposed to their behavioural dispositions, "is both mediated by symbols and dependent on many of the same mental operations and neural substrates as are critical to symbolic activities" (p. 427). He concludes that human mindreading abilities therefore developed *after* the acquisition of symbolic communication; the necessary prefrontal cortex development being the result of selection for the ability to use a symbolic communication system (pp. 427-8).

This sequence is not necessarily the case, however. As I see it, the order would be reversed, and in this section, I'll explain why I think the order of the evolution of these abilities and practices must have been this way. A brief summary first, and then some details: In a complex norm-abiding society that employs ritualized punishment for transgression of the social order (like chimpanzee societies, and like our Australopithecine ancestors' societies are thought to have been), tacit ground-level mindreading abilities could easily develop; as I've already explained. This situation gave rise to the ability to attribute an inter-related system of intentional states. Once such abilities spread throughout a population, the possibility of signalling (initially gesturally, perhaps) with the intention of getting others to attribute reasons for your act of signalling, is possible. This created the conditions under which symbolic human languages, as norm-governed uses of signs used to perform speech acts, could arise. These languages reinforced and developed the developing folk psychology, while at the same time developed a shared norm-governed *practice* (as I described at the end of Chapter Five) interrelating the folk psychology with people's behavior. The norms make particular contentful intentional states appropriate to attribute to others, and entitle people to expect those others to act as one with the intentional states attributed should act.

That's the overview. Now for some details.[136]

It seems likely that the insight into the utility of "labelling" other creatures as having intentional states (ground-level mindreading) was the insight that is a precondition for the first symbolic *system*. The recognition of a single abstract intentional state (e.g. X knows that the food box is empty) could certainly occur in a creature that had not yet developed a language in which to explicitly express this recognition.[137]. All that would be required would be: (1) a complex social and political environment in which there is plenty of opportunity to observe other's behaviour, (2) enough intelligence and memory to recognize the complex pattern of cause and effect associations, and (3) at least some abstraction abilities, to posit an intervening variable which might simplify the pattern recognized.

Our Australopithecine ancestors probably had at least some capacity for abstraction. As Deacon points out, the prefrontal cortex is the area of the brain employed in social interactions and in abstract thought. For example, damage to the prefrontal cortex causes defects in the ability to appreciate others' perspectives (autism), and exaggerated prefrontal activity appears to be at least partly responsible for the hypersociality of people with Williams' syndrome.[138] This suggests, then, that in relatively intelligent but non-linguistic species whose members live in very complex social and political groups with extensive system of norms governing what one should do and punishments for transgressions, such as those our distant ancestors probably lived in, selection in favour of those who are more socially and politically astute (manipulative, able to appreciate what they can "get away with", etc.) could effect some prefrontal cortex development. Even if Deacon is correct in supposing that some amount of prefrontal cortex development would be necessary for the ability to attribute such an abstract entity as an intentional state to another, this does not mean that this could only happen after the acquisition of symbolic communication.

[136] The following summary perhaps goes rather too quickly over some of these steps. Each could easily ba a paper by themselves. Many of them have been.

[137] Both Andrew Whiten (1996b, p. 288) and Jonathan Bennett (1976, p. 110) argue that a relatively intelligent but languageless creature could come to have beliefs about others' intentional states.

[138] Deacon discusses extensively the role of the prefrontal cortex both in cases of autism, where there is decreased activity in the prefrontal cortex, and with Williams' syndrome, where prefrontal cortical activity is abnormally high (pp. 264-78).

This ground-level ability to tacitly attribute a single intentional state would be prior to the adoption of any symbolic system. The initial recognition of one intentional state, would have quickly been followed by reapplication of this tactic to attribute more such abstract intervening variab les. The ability to label others as having many different intentional states and to explain and predict their behaviour would give its possessors a significant selective advantage. Selection for this abstract cognitive skill would probably bring about further development in the prefrontal cortex and thus even more astute abstraction abilities.

It's at this stage that full-blown mindreading, human languages, and the practices that integrate them could develop together. I'll discuss these in turn, as discrete steps, even though they really feed one another's evolution. The initial steps would be very small —though significant— ones, but the combination "bootstraps" the effect, so that much progress could be made rather quickly (evolutionarily speaking).

It would not be long before the abstract labels one applies to fellow creatures could themselves be thought about abstractly, and connections between the labels themselves could be noticed. The labels could be seen to relate to one another as well as to the events that are predictive of them and the behaviours that they predict. For instance, noticing how the label *X wants meat,* and *X knows I have meat,* relate to and predict the label, *X intends to get my meat* would simply require a recursive application of the same kind of abstraction abilities. The discovery of this combinatorial system, in which beliefs relate to desires, each relate to intentions, some types of beliefs relate to others and so on, would be enable even more effective prediction of others' behaviour; an even greater advantage in the probably now increasingly complex social and political environment. The resulting system of attributed intentional states which as general types relate to one another in various ways, and which are also related to (abstract) aspects of the world that are themselves held to be related to one another, seems quite close to the description Deacon gives (pp. 79-101) of a symbolic *system*.

Thus the first symbolic system would not, as Deacon argues, be a symbolic system for *communication*, but a symbolic *representational* system, in Deacon's sense (see pp. 99-100). It's the beginning of a folk psychology; a system for representing the many different intentional states one attributes to others and the ways these intentional states related to one another and to behaviour. Andrew Whiten (1993, p. 38708) makes a similar remark, to the effect that a theory of mind should be based in the ability to integrate judgements about others' mental states. Otherwise it's just a *hypothesis* about a mental state, not a *theory* of mind.

This might be a symbolic system by Deacon's definition, but it's not yet one by Peirce's. A shared normative dimension is lacking. The above system could be discovered and employed by *one individual*, as their own personal theory about other folks' mental lives. For this reason, it also fits Dennett's description of a creature that can adopt the intentional stance too: one attributes intentional state to others, and the reason for doing so, and the justification[139] for continuing to do so, is the predictive and explanatory success it affords.

However, this next step in the evolutionary and cultural progression I'm tracing, comes when we consider the ways that these abilities play themselves out by being possessed by many individuals in the community. The above kinds of deceptive uses of signals, and the uses of these mindreading skills as a defense against such deception bring about a mindreading "arms race" mean that it wouldn't be long before either many people learned this trick of adopting the abstract system of intentional states, or there was selection for the ability to do this.

This shows, furthermore, how sophisticated human linguistic interactions of the sort I've been talking about could have developed. Once mindreading abilities were common in a population, and in use alongside (i.e.. to defend against) manipulative signalling behaviour, this combination "bootstrapped" signalling interactions to more sophisticated levels. This "leap" happened because in addition to attributing intentional states to others, people also know that others can attribute intentional states to them. Thus people become able to take advantage of others' ability to attribute aims and beliefs to them, by performing an action with the intention of getting the other to attribute a certain intentional state to them as a reason for their so acting.

And once this kind of action becomes commonplace, the intention to get the other to attribute that reason can likewise become apparent, through the shared knowledge that that kind of action is often used for that kind of reason.

[139] Since it's a theory employed by one person alone, there is no difference between the theory being applied incorrectly and the theory itself being incorrect. Here, as Dennett argues, we can't talk about correctness. The only standard the theory can be assessed by is its predictive success and the pragmatic justification such successes confer on the theory.

The action gets treated as a signal, whose "object" is (in part) a certain reason for signalling (e.g. my intention to get you to help me in some way).

The fact that the previously manipulative motivations for signalling becomes recognizable, creates conditions under which a less ostensibly manipulative, more cooperative, social environment could develop. The use of signals to induce others to do as explicitly I intend they do, creates conditions whereby a norm of working together, helping one another, could be encouraged, and even enforced. This appears to have happened. Christopher Boehm (1997b) reports the well-supported thesis that "our prehistoric human predecessors remained consistently egalitarian for scores, probably hundreds, of millennia" (p. S101). Such egalitarian societies, says Boehm, "arrive at very general covenants about how people should behave: they favour political autonomy, sharing, cooperation, and being helpful to others" (p104). These norms are rigidly enforced. People trying to assert dominance, secure resources for themselves, and break this cooperative pattern are subject to serious sanctions "that include not only gossip, but direct criticism, ridicule, ostracism, exile, and execution" (p. S104) (see also Boehm 1997a).

In such a society, where people generally attribute intentional states to others, and where norms of cooperative behaviour are enforced, signallers could signal with the sort of higher-order self-referential intentions that I based linguistic interactions upon in Chapter Four. I can aim, through signalling, to get others to react, not to my signal itself, but to *the intention with which I signal.* I can signal with the *expectation* that the person I signal to will be able to discern the reason for my signalling so in my intention to elicit a particular response from them. I can also expect that it's likely that they will so respond (especially because following the norm of not being coercive and manipulative is expected of me too). If I want to inform someone of the location of ripe berries, I can perform some kind of pantomime, for instance, of the motions of eating berries with a "Delicious!" expression on my face, and point in the direction to go. And I can do this with the expectation that they can deduce my reason for signalling: to indicate that there are *edible berries* in that direction, in order to help them find the berries. The pantomime part of the sign is iconic in form, but is also functions symbolically, in that the person must recognize which aspects of my action are relevant, and interpret them to be referring to *edible berries* thataway.

And thus we can begin to get conventions of sign-interpretation developing. As people use signs to induce recognition of the intention to get others to respond in certain ways, some signs will spread. Particularly easily interpretable signs would be picked up and used often, once shown to a few others. These signs could very well be largely gestural rather than vocal at first, given the ease of interpretation of the reference of certain iconic gestures (but still interpreted with an eye to the signallers reasons for signalling).[140] Such symbols, furthermore, become common currency. And once they do, their form can get shortened, and be abstracted away from their iconic origins, into something more arbitrary that operates purely according to norms for how it should be (is *norm*ally) used.

This situation encourages further escalation of the mindreading "arms race". Here, even more sophisticated manipulation can take place, since I can signal, not just intending to get you to do something, but intending to manipulate your beliefs about my intentions. They person I informed about the berries, for instance, also could be concerned about my motives friendly or deceitful. I could be signalling to get them to believe that I'm being friendly and helping them, when in fact my true intention is just to get them to go away, so I can have the berries right behind me all to myself. Of course if the norm for not doing this kind of thing is strictly enforced, I need even more sophisticated mindreading and abstraction abilities to ensure that I do everything necessary to get away with employing such tactics.

Thus the normative dimension of sign-use and mindreading really comes into play. This is especially so when people fully appreciate the fact that others attribute intentional states to you, based on how you behave. A concern for the content of the states that others attribute to you, and the disinclination to give others reason to accuse you of being deceitful, creates conditions whereby people start to use public signs to refer to intentional states themselves. By employing public signs to clarify the precise nature of the reason for your action (perhaps to prevent a beating after being accused of deceptive intentions), people are able to *explicitly* characterize the content of their reasons, and the reasons they attribute to others. Because of the ability to use public signs to make one's intentions explicit, and the norm-governed practice that would already exist for the proper uses of such signs, the content of the intentional

[140] Armstrong's (1998) review of Deacon's book is largely focused on this theory that gestures could well have been the first symbolic signs. U. T. Place (2000) similarly summarizes eleven pieces of evidence for the view that vocal language must have been preceded by an earlier stage of gesture. Place argues for a similar progression from the iconic to the symbolic nature of gestural language, with vocalizations being a later entry.

states attributed to people gets drawn out of personal "theories of mind" and into a shared normative practice of attributing intentional states. Disputes will arise, for instance, based in differences between the content of the intention that I attributed to you, and the content of your intention *as you understood that intention*. Such disputes and the need to clarify differences, requires norms of how words should be used to attribute contentful, explicit, intentional states to people. The need to clarify differences between the content of intentional states entails the need for *shared* norms for how words should be used to make such contents explicit.

This kind of social and cognitive environment enables the normative practices I talked about at the end of Chapter Five to develop and be preserved and enforced. For instance, labelling and keeping track of the trustworthiness of certain neighbours would be a distinct advantage. Such an assessment would be based on the connections between the intentional states people attribute to the individual in question, and the individual's subsequent behaviour. Someone who does that kind of thing should be labeled with such and so contentful intentional state, and someone with such and so intentional state should behave in so and so ways. When the individual doesn't behave in ways that the intentional state appropriately attributed to then indicates that they should behave, trouble is afoot. Retribution could even be warranted if the person behaved deceitfully in behaving such that the contentful intention they knew others would describe them as having was different from the one that they would describe themselves as having.

Thus norms develop, governing exactly what public symbols should be used to describe the content of intentional states attributed based on particular behaviour, and norms governing the specific kinds of behaviour one expects of someone whose intentional states are described using those public symbols. People who wish to avoid accusations of deceitfulness, and who wish to conform to the conventions of cooperation and helpfulness, should behave according to the norms for behaviour of people with their particular intentional states. But these are not the intentional states that they attribute to themselves, but the intentional states that others are licensed to attribute to them.

Thus self-consciousness also emerges within this mindreading and linguistic development; here self-consciousness is the ability to keep track of the intentional states that fit *norm*ally with the actions one performs, and with the actions one is counterfactually disposed to perform.

6.9 *Summary and conclusions*

This completes my quick survey of the evolutionary emergence of the kind of account of intentionality that I have been building up throughout this work. This account, recall, culminated in the norm-based practices of attributing intentionality (goal-directedness) to people's actions, and thus to people based on the goals that count as reasons for those actions. This is complemented by the way these norms constrain people's behaviour, by obliging in ways consistent with the intentional states attributed to them.

Chapter One laid out the nature of an ecological, embodied action approach to explaining how it is that human beings can do all the remarkable things we can do. Chapter Two laid out the problem of intentionality, as embodied approaches to cognition could conceive of it, and the problem of naturalizing intentionality as it commonly features in philosophy of mind. The subsequent three chapters elaborated and detailed and defended that account of intentionality and the way it is instituted in shared norm-governed practices; particularly linguistic ones, although it depends on a large foundation of non-linguistic, non-intentional tacit norms and practices. I have argued that intentionality is an institutional property of people's actions (the goal the action counts as being directed towards), and derivatively of people's intentional states, states attributed to them as reasons for their actions. The neurological processes that cause these actions to take place also only have intentionality derivatively, by virtue of abstract generalizations about the actions each enables and the intentionality the public practices confer on such actions.

This last chapter has shown how the problem of naturalizing intentionality gets reconfigured, and solved, in this ecological embodied approach to cognition. Most theorists recommend the tactic of reducing the intentionality attributed to a neurological representation to purely physical, causal properties of that representation (and perhaps of the system it functions within). This tactic is impossible to employ successfully, however, because it expects what turns out to be a normative property to be *justified* in physical terms. The way to give a naturalistic explanation for intentionality is to embrace this institutional normative foundation, and to explain the evolutionary history of the emergence of the practices and actions within which intentionality is instituted. Thus explanations of intentionality by appeal to these practices and their norms have a respectable naturalistic heritage.

Thus, I can conclude that the embodied approach to cognition, thought of in this way, as explaining the capacities of embodied, socially situated agents, has a distinct advantage. We might say that a group of cognitive scientists practicing this form of analysis will have a selective advantage over a group of cognitive scientists practicing traditional brain-centered, individualist, conceptions of human cognitive systems.

Of course, they don't *have* to be competitive, nor independent. We might also say that each group could learn some of the others' analytical tools and methodological techniques, and develop an ecological, embodied approach with a lot to say about how the brain's operations contribute to the capacities of human cognitive agents. This seems to be the most promising direction for the future of cognitive science.

Bibliography

Abraham, Ralph H. and Christopher D. Shaw. 1992. *Dynamics—the Geometry of Behavior*. Santa Cruz, CA: Addison-Wesley.

Alston, William P. 1964. *Philosophy of Language*. Englewood Cliffs, NJ: Prentice-hall.

Armstrong, David F. 1998. Review of *The Symbolic Species: The Co-evolution of Language and the Brain*. *Evolution of Communication* 2(1): 161-9.

Austin, John Langshaw. 1961. A Plea for Excuses. In *Philosophical Papers by the late J. L. Austin* . Oxford: At The Clarendon Press. 123-152. (Originally published in *The proceedings of the Aristotelian Society*, 1956)

Austin, John Langshaw. 1962. *How to Do Things With Words*. Oxford: Clarendon Press. (William James lectures. 1955)

Baldwin, James Mark. 1895. Consciousness and Evolution. *Science* 2: 219-23.

Baldwin, James Mark. 1896. On Criticisms of Organic Selection. *Science* 4: 727.

Baldwin, James Mark. 1902. *Development and Evolution*. New York: Macmillan.

Barkow, Jerome H., Leda Cosmides and John Tooby, Eds. 1992. *The Adapted Mind : Evolutionary Psychology and the Generation of Culture*. New York,: Oxford University Press.

Baron-Cohen, Simon, Helen Tager-Flusberg and Donald J. Cohen, Eds. 1993. *Understanding Other Minds: Perspectives from Autism*. Oxford: Oxford University Press.

Bechtel, William. 1990. Multiple Levels of Inquiry in Cognitive Science. *Psychological Research* 52: 271-281.

Bechtel, William and Adele Abrahamsen. 1991. *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Cambridge, Mass.: Basil Blackwell.

Bennett, Jonathan. 1976. *Linguistic Behaviour*. Cambridge: Cambridge University Press.

Bestor, Thomas Wheaton. 1976. Dualism and Bodily Movements. *Inquiry* 19: 1-26.

Bestor, Thomas Wheaton. 1979. Gilbert Ryle and the Adverbial Theory of Mind. *The Personalist* 60(June): 233-42.

Boehm, Christopher. 1997a. Egalitarian Behaviour and the Evolution of Political Intelligence. In *Machiavellian Intelligence II: Extensions and Evaluations* Eds. Andrew Whiten and Richard W. Byrne. Cambridge: Cambridge University Press. 144-173.

Boehm, Christopher. 1997b. Impact of the Human Egalitarian Syndrome on Darwinian Selection Mechanics. *The American Naturalist* 150(Supplement): S100-121.

Boesch, C. and H. Boesch. 1990. Tool Use and Tool Making in Wild Chimpanzees. *Folio Primatologica* 54: 86-99.

Boyd, Robert and Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, Robert and Peter J. Richerson. 1992. Punishment allows the Evolution of Cooperation (or Anything Else) in Sizable Groups. *Ethology and Sociobiology* 13: 171-95.

Braitenberg, Valentino. 1984. *Vehicles.* Cambridge, Mass: MIT Press, A Bradford Book.

Brandom, Robert B. 1994. *Making it Explicit: Reasoning, Representing and Discursive Commitment.* Cambridge, MA: Harvard University Press.

Brentano, Franz. 1874/1973. The Distinction Between Mental and Physical Phenomena. In *Psychology from an Empirical Standpoint* . New York: Humanities Press. (A. Rancurello, D. B. Terrell and L. McAlister transl.) (Translation of *Psychologie vom Empirischen Standpunkt*)

Bretherton, I., S. McNew and M. Beeghly-Smith. 1981. Early person knowledge as expressed in gestural and verbal communication: When do infants acquire a theory of mind? In *Social Cognition in Infancy* Eds. M. Lamb and L. Sherrod. Hillsdale, NJ: Erlbaum.

Brewer, S. M. and W. C. McGrew. 1990. Chimpanzee use of a Tool-set to get Honey. *Folio Primatologica* 54: 100-104.

Brooks, Rodney and Anita M. Flynn. 1989. Fast, Cheap and Out of Control: A Robot Invasion of the Solar System. *Journal of the British Interplanetary Society* 42: 478-85. (Available from http://www.ai.mit.edu/people/brooks/brooks.html)

Brooks, Rodney and P. Maes, Eds. 1994. *Artificial Life.* Cambridge, Mass.: MIT Press.

Brooks, Rodney A. 1991a. *Intelligence Without Reason.* MIT AI Lab Memo 1293. April 1991. (Available from: http://www.ai.mit.edu/people/brooks/papers.html)

Brooks, Rodney A. 1991b. Intelligence without Representation. *Artificial Intelligence* 47: 139-159.

Brooks, Rodney A. and Lynn Andrea Stein. 1993. *Building Brains for Bodies.* MIT AI Lab Memo 1439. August 1993. (Available from: http://www.ai.mit.edu/people/brooks/papers.html)

Bruner, Jerome and Carol Feldman. 1993. Theories of Mind and the Problem of Autism. In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press. 267-291.

Buber, Martin. 1923/1970. *I and Thou.* Walter Kaufmann trans. New York: Charles Scribner's Sons.

Button, Graham, Jeff Coulter, John R. E. Lee and Wes Sharrock. 1995. *Computers, Minds and Conduct.* Cambridge, UK: Polity Press.

Carroll, Lewis. 1871. *Through the Looking Glass.* London: Macmillan.

Carruthers, Peter. 1996. Autism as mind-blindness: an elaboration and defence. In *Theories of Theories of Mind* Eds. P. Carruthers and P. K. Smith. Cambridge: Cambridge University Press. 257-273.

Chalmers, David. 1996. *The Conscious Mind.* Oxford University Press.

Chandler, M., A. S. Fritz and S. Hala. 1989. Small Scale Deceit: Deception as a Marker of 2- 3- and 4-year old's Early Theories of Mind. *Child Development* 60: 1263-77.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin and Use.* New York: Praeger.

Chomsky, Noam. 1988. *Language and Problems of Knowledge: The Managua Lectures.* Cambridge, Mass.: MIT Press.

Clancey, William J. 1997. *Situated Cognition : On Human Knowledge and Computer Representations.* Cambridge, U.K.: Cambridge University Press.

Clark, Andy. 1998. *Being There: Putting Brain, Body and World Together Again.* Cambridge, Mass: MIT Press, A Bradford Book.

Clark, Andy and Josefa Toribio. 1994. Doing Without Representing? *Synthese* 101: 401-31.

Cook, John. 1969. Human Beings. In *Studies in the Philosophy of Wittgenstein* Ed. Peter Winch. Routledge and Kegan Paul. 117-151. (Reprinted in John V. Canfield (ed.) *The Philosophy of Wittgenstein* (A fifteen volume set). Volume 12, "Persons": 59-93.)

Cummins, Robert. 1989. *Meaning and Mental Representation.* Cambridge, Mass: MIT Press. A Bradford Book.

Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason and the Human Brain.* New York: Avon Books.

Davidson, Donald. 1963. Actions, Reasons, and Causes. *Journal of Philosophy* 60: 685-700. (Reprinted in Davidson (1980) *Essays on Actions and Events.* Clarendon Press, Oxford: 3-19.)

Davidson, Donald. 1980. Mental Events. In *Essays on Actions and Events* . Clarendon Press: Oxford. 3-19.

Davidson, Donald. 1986. A Nice Derangement of Epitaphs. In *Truth and Interpretation* Ed. Ernest LePore. Basil Blackwell. 433-446.

Dawkins, Richard. 1976. *The Selfish Gene.* Oxford: Oxford University Press.

Dawkins, Richard and John R. Krebs. 1978. Animal Signals: Information or Manipulation? In *Behavioural Ecology: An Evolutionary Approach* Eds. J. R. Krebs and N. B. Davies. Oxford: Blackwell. 282-309.

Dawson, Michael R. W. 1998. *Understanding Cognitive Science.* Malden, Mass.: Blackwell.

Deacon, Terrance W. 1997. *The Symbolic Species: The Co-evolution of Language and the Brain.* New York: W. W. Norton & Co.

Dennett, Daniel. 1971. Intentional Systems. *Journal of Philosophy* 68(4): 87-106. (Reprinted in Dennett's (1978) *Brainstorms* Montgomery, VT: Bradford Books.)

Dennett, Daniel. 1987. *The Intentional Stance.* Cambridge, Mass.: MIT Press, A Bradford Book.

Dennett, Daniel. 1991a. Real Patterns. *Journal of Philosophy* LXXXVIII(1): 27-51.

Dennett, Daniel. 1995a. *Darwin's Dangerous Idea.* New York: Simon and Schuster.

Dennett, Daniel. 1995b. The Unimagined Preposterousness of Zombies. *Journal of Consciousness Studies* 2: 322-26.

Dennett, Daniel. 1996. *Kinds of Minds.* New York: Basic Books.

Dennett, Daniel. 1997. Making Tools for Thinking. *Paper presented at Metarepresentation, SFU 10th Annual Cognitive Science Conference,* Simon Fraser University, Vancouver, BC. Feb 7-8, Dennett, Daniel C. 1978. The Abilities of Men and Machines. In *Brainstorms* . Cambridge, MA: MIT Press: A Bradford Book.

Dennett, Daniel C. 1991b. *Consciousness Explained*. Boston: Little, Brown.

Dennett, Daniel C. 1994. The Practical Requirements for Making a Conscious Robot. *Philosophical Transactions of the Royal Society* A(349): 133-46. (Also available at: http://www.tufts.edu/as/cogstud/papers/practic.htm)

Descartes, René. 1911. Reply to Objections V. In *The Philosophical Works of Descartes* Eds. E. S. Haldane and G. R. T. Ross. New York: Dover. (2 of 2)

Donnellan, Keith. 1969. Putting Humpty Dumpty Together Again. *The Philosophical Review* 77:

Dreyfus, Hubert. 1972. *What Computers Can't Do*. Cambridge, MA: MIT Press. (Second (1979) edition, with new preface.)

Dreyfus, Hubert. 1991. *Being-in-the-world: A Commentary on Heidegger's Being and Time*. Harper & Row.

Dreyfus, Hubert. 1992. *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

Dreyfus, Hubert and Stuart Dreyfus. 1982. *Mind Over Machine*. Glencoe, IL: Free Press.

Dreyfus, Hubert L. 1996. The Current Relevance of Merleau-Ponty's Phenomenology of Embodiment. *The Electronic Journal of Analytic Philosophy* 4(spring): (EJAP is at: http://www.phil.indiana.edu/ejap/)

Easton, Susan M. 1978. Conventionalism and the limits to social change. *Social Praxis* 5(3-4): 323-41.

Ebersole, Frank. 1967. Where the Action Is. In *Things We Know: Fourteen Essays on Problems of Knowledge* . University of Oregon Press.

Elitzur, A. 1989. Consciousness and the Incompleteness of the Physical Explanation of Behavior. *Journal of Mind and Behavior* 10: 1-20.

Fenton, Andrew. 2000. Human Knowledge as Animal Knowledge: Broadening the Community of Knowers. *Canadian Philosophical Association Conference*, Edmonton, Alberta,

Fodor, Jerry. 1981. The Mind-Body Problem. *Scientific American* 244(January): 114-22.

Fodor, Jerry. 1984. Semantics, Wisconsin style. *Synthese* 59(231-250): (Reprinted in Fodor 1990. *A Theory of Content* MIT Press: 31-49)

Fodor, Jerry. 1987. Meaning and the World Order. In *Psychosemantics* . Cambridge Mass.: MIT Press. A Bradford Book. 97-133.

Fodor, Jerry. 1990. A Theory of Content I: The Problem. In *A Theory of Content and Other Essays* . Cambridge, Massachusetts: MIT Press. A Bradford Book. 51-88.

Fodor, Jerry and Zenon Pylyshyn. 1988. Connectionism and Cognitive Architecture: a Critical Analysis. *Cognition* 28: 3-71.

Foss, Jeffrey E. 1995. Materialism, Reduction, Replacement, and the Place of Consciousness in Science. *Journal of Philosophy* XCII(8, August): 401-29.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin Company.

Gould, Stephen Jay. 1989. Tires to Sandals. *Natural History* (April): 8-15.

Grice, H. P. 1957. Meaning. In *Philosophical Logic* Ed. P. F. Strawson. Oxford: Oxford University Press. (Originally published in *Philosophical Review* 66 (1956), 377-388)

Grice, H. P. 1975. Logic and Conversation. In *The Logic of Grammar* Eds. Donald Davidson and Gilbert Harman. Encino and Belmont, CA: Dickenson Publishing Company. 64-75. (This is from Grice's William James Lectures, Delivered at Harvard University in 1967))

Grice, H. P. 1989. *Studies in the Way of Words*. Cambridge, Mass: Harvard University Press.

Hannah, A. C. and W. C. McGrew. 1987. Chimpanzees Using Stones to Crack Open Oil Palm Nuts in Liberia. *Primatea* 28: 31-46.

Haugeland, John. 1982. Heidegger on Being a Person. *Nous* XVI(1): 15-26.

Haugeland, J. 1990. The Intentionality All-Stars. *Philosophical Perspectives* 4: 383-427.

Haugeland, John. 1991. Representational Genera. In *Philosophy and Connectionist Theory* Eds. William Ramsey, Stephen Stich and David Rumelhart. New Jersey: Erlbaum. 61-90.

Haugeland, John. 1995. Mind Embodied and Embedded. In *Mind and Cognition* Eds. Y.-H. Houng and J.-C. Ho. Taipei: Academia Sinica.

Hauser, Marc. 1996. *The Evolution of Communication*. Cambridge, MA: MIT Press: a Bradford Book.

Heidegger, Martin. 1927/1962. *Being and Time*. John Macquarrie and Edward Robinson trans. SCM Press.

Hofstadter, Douglas R. 1979. *Gödel, Escher, Bach : An Eternal Golden Braid*. New York: Basic Books.

Hofstadter, Douglas R. and Daniel C. Dennett, Eds. 1981. *The Mind's I: Fantasies and Reflections on Self and Soul*. New York: Basic Books.

Hume, David. 1739. *A Treatise on Human Nature, Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects, and Dialogues Concerning Natural Religion*. 1874 edn. London: Longman's Green. (edited by T.H. Green and T. H. Grose)

Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press, A Bradford Book.

Jackson, Frank. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32: 127-136.

Kirk, Robert. 1974. Sentience and behaviour. *Mind* 81(43-60):

Kirshner, David and James A. Whitson, Eds. 1997. *Situated Cognition: Social, Semiotic, and Psychological Perspectives*. Mahwah, NJ: Erlbaum.

Krebs, John R. and Richard Dawkins. 1984. Animal Signals: Mind Reading and Manipulation. In *Behavioural Ecology: An Evolutionary Approach* Eds. J. R. Krebs and N. B. Davies. Oxford: Blackwell. (Second edn.) 380-401.

Lakatos, Imre. 1974. Falsification and the Methodology of Scientific Research Programmes. In *Criticism and the Growth of Knowledge* Eds. I. Lakatos and A. Musgrave. Cambridge University Press:

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, George and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* New York: Basic Books.

Langton, Christopher. 1995. *Artificial Life: An Overview.* Cambridge, MA: MT Press, A Bradford Book.

Leslie, Alan. 1991. Theory of Mind Impairment in Autism: Evidence for a Modular Mechanism of Development? In *Natural Theories of Mind* Ed. Andrew Whiten. Oxford: Basil Blackwell.

Leslie, Alan and Daniel Roth. 1993. What Autism Teaches us about Metarepresentation. In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press. 83-111.

Long, Douglas C. 1964. The Philosophical Concept of a Human Body. *Philosophical Review* LXXIII(July): 321-337.

Loveland, Katherine and Belgin Tunali. 1993. Narrative language in Autism and the Theory of Mind Hypothesis: a Wider Perspective. In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press. 245-66.

MacIntyre, Alisdair. 1981. *After Virtue: A Study in Moral Theory.* London: Duckworth.

Malcolm, Norman. 1978. Thinking. In *Wittgenstein and His Impact on Contemporary Thought (Proceedings of the second International Wittgenstein Symposium, 1977)* Eds. Elisabeth Leinfellner and Werner Leinfellner. 411-419.

Masur, E. F. 1983. Gestural Development, dual-directional signalling, and the Transition to Words. *Journal of Psycholinguistic Research* 12: 93-109.

Maturana, Humberto, Jorge Mpodozis and Juan Carlos Letelier. 1995. Brain, Language and the Origin of Human Mental Functions. *Biological Research* 28: 15-26. (AS reprinted at http://www.informatik.umu.se/~rwhit/Mat&Mpo&Let(1995).html)

Maturana, Humberto R. 1975. The organization of the living: A theory of the living organization. *International Journal of Man-Machine Studies* 7: 313-332.

Maturana, Humberto R. and Francisco Varela. 1980. *Autopoiesis and cognition: the realization of the living.* Dordrecht, Holland: D. Reidel Publishing Co.

McGrew, W. C. and M. E. Rogers. 1983. Chimpanzees, tools, and termites: New Record from Gabon. *American Journal of Primatology* 5: 171-4.

Meltzoff, Andrew. 1995. Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children. *Developmental Psychology* 31(May): 813-850.

Meltzoff, Andrew and Alison Gopnik. 1993. The Role of Imitation in Understanding Persons and Developing a Theory of Mind. In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press. 335-366.

Meltzoff, Andrew and M. K. Moore. 1977. Imitation of Facial and Manual Gestures by Human Neonates. *Science* 198: 75-78.

Meltzoff, Andrew and M. K. Moore. 1983. Newborn Infants Imitate Adult Facial Gestures. *Child Development* 54: 702-9.

Merleau-Ponty, Maurice. 1942. *The Structure of Behavior*. Beacon. (Translation of *La Structure du Comportment*, Presses Universites de France, 1963)

Merleau-Ponty, Maurice. 1945. *Phenomenology of Perception*. Routledge and Kegan Paul (1962).

Millikan, Ruth. 1989. Biosemantics. *Journal of Philosophy* 86(6): (Reprinted in Stephen Stich and Ted Warfield eds. 1994. *Mental Representation: A Reader.* Oxford: Blackwell.)

Moody, Todd. 1994. Conversations with zombies. *Journal of Consciousness Studies* 1: 196-200.

Morgan, Elaine. 1995. *The Descent of the Child : Human Evolution From a New Perspective*. New York: Oxford University Press.

Nagel, Thomas. 1979. What Is It Like to Be a Bat? *Philosophical Review* 83: 435-450.

Neisser, Ulric. 1989. Direct Perception and Recognition as Distinct Perceptual Systems. *Eleventh Annual Meeting of the Cognitive Science Society,* Ann Arbor, MI, (As cited in William Bechtel and Adele Abrahamsen. 1991. *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Cambridge, Mass.: Basil Blackwell.)

Ó Nualláin, Sean. 1995. *The Search For the Mind: A New Foundation for Cognitive Science*. Norwood, NJ: Ablex Publishing Company.

Okrent, Mark. 1996. Why the Mind Isn't a Program (But Some Digital Computer Might Have a Mind). *The Electronic Journal of Analytic Philosophy* 4(spring): (EJAP is at: http://www.phil.indiana.edu/ejap/)

Peirce, Charles Sanders. 1960. *Collected Papers*. Cambridge, MA: The Belknap Press of Harvard University Press. (Eds. Charles Hartshorne and Paul Weiss)

Pfeifer, Rolf and Christian Scheier. 1999. *Understanding Intelligence*. Cambridge, MA: MIT Press.

Piaget, Jean. 1954. *The Construction of Reality in the Child.* New York: Basic Books.

Pinker, Stephen. 1994. *The Language Instinct: How the Mind Creates Language*. New York: William Morrow.

Pinker, Stephen and Paul Bloom. 1990. Natural Language and Natural Selection. *Brain and Behavioural Sciences* 13: 707-784.

Place, U. T. 2000. The Role of the Hand in the Evolution of Language. *Psycholoquy* 11(007): (Psycholoquy is an on-line journal, available at: http://www.princeton.edu/~harnad/psyc.html)

Polanyi, Michael. 1958. *Personal Knowledge: Towards a Post-Critical Philosophy.* Chicago: University of Chicago Press.

Port, Robert F. and Timothy van Gelder. 1995a. It's About Time: An Overview of the Dynamical Approach to Cognition. In *Mind as Motion: Explorations in the Dynamics of Cognition* Eds. Robert F. Port and Timothy van Gelder. Cambridge, Mass: MIT Press.

Port, Robert F. and Timothy van Gelder, Eds. 1995b. *Mind as Motion:*. Cambridge, Mass: MIT Press.

Putnam, Hilary. 1975a. The Meaning of Meaning. In *Minnesota Studies in the Philosophy of Science* Ed. Keith Gunderson. Minneapolis: University of Minnesota Press. 131-193.

Putnam, Hilary. 1975b. The Nature of Mental States. In *Mind, Language and Reality: Philosophical Papers* . Cambridge: Cambridge University Press. (2 of 429-440.

Putnam, Hilary. 1981. *Reason, Truth, and History*. Cambridge: Cambridge University Press.

Reddy, V. 1991. Playing with Others' Expectations: Teasing and Mucking about in the First Year. In *Natural Theories of Mind* Ed. Andrew Whiten. Oxford: Basil Blackwell.

Roberts, Monty. 1997. *The Man who Listens to Horses*. New York: Random House.

Rutter, Michael. 1991. Autism as a Genetic Disorder. In *The New Genetics of Mental Illness* Eds. P. McGuffin and R. Murray. Oxford: Heinemann Medical.

Rutter, Michael and Anthony Bailey. 1993. Thinking and Relationships: Mind and Brain (Some Reflections on Theory of Mind and Autism). In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press. 481-504.

Ryle, Gilbert. 1949. *The Concept of Mind*. Hammondsworth, UK: Penguin Books.

Sacks, Oliver. 1996. *An Anthropologist on Mars*. Toronto: Vintage Canada.

Sanders, John T. 1996. An Ecological Approach to Cognitive Science. *The Electronic Journal of Analytic Philosophy* 4(spring): (EJAP is at: http://www.phil.indiana.edu/ejap/)

Schiffer, Stephen R. 1972. *Meaning*. Oxford: Clarendon Press.

Searle, John. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Oxford: Oxford University Press.

Searle, John. 1980. Minds, Brains and Programs. *The Behavioral and Brain Sciences* 3: 417-424.

Searle, John. 1987. Indeterminacy, Empiricism, and the First Person. *Journal of Philosophy* LXXXIV(3, March):

Searle, John. 1990. Is the brain's mind a computer program? *Scientific American* 262(1): 26-31.

Searle, John. 1994a. The Connection Principle and the Ontology of the Unconscious: A Reply to Fodor and Lepore. *Philosophy and Phenomenological Research* 54: 847-55.

Searle, John. 1994b. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press, A Bradford Book.

Searle, John R. 1995. *The Construction of Social Reality*. New York: Free Press.

Shannon, Benny. 1993. *The Representational and the Presentational*. Hemel Hempstead: Harvester Wheatsheaf.

Simon, H.A. and A. Newell. 1958. Heuristic problem solving: The next advance in operations research. *Operations Research* 6:

Skarda, C. and W. Freeman. 1987. How Brains Make Chaos in Order to Make Sense of the World. *Behavioural and Brain Sciences* 10: 161-95.

Smith, Peter and O. R. Jones. 1986. *The Philosophy of Mind: An Introduction*. London: Cambridge University Press.

Smolensky, P. 1988. On the Proper Treatment of Connectionism. *Behavioural and Brain Sciences* 11(1-74):

Sneddon, Andrew. forthcoming. Naturalistic Study of Culture. *Mind, Culture and Activity*

Sober, Elliot. 1984. Holism, Individualism, and the Units of Selection. In *Conceptual issues in Evolutionary Biology* Ed. Elliott Sober. Cambridge, Mass.: MIT Press. 184-299.

Sober, Elliot. 1991. Models of Cultural Evolution. In *Essays in the Philosophy of Biology* Ed. P Griffiths. Kluwer. 478-92. (As reprinted in *Conceptual Issues in Evolutionary Biology* Ed. Elliott Sober. Cambridge, Mass.: MIT Press. (Second, 1994 edn.) 478-92.)

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach.* Cambridge, Mass: Blackwell.

Sprague, Elmer. 1999. *Persons and Their Minds: A Philosophical Investigation.* Boulder, CO: Westview Press.

Sterelny, Kim. 1990. *The Representational Theory of the Mind.* Oxford: Basil Blackwell.

Stillings, N. , M.H. Feinstein, J.L. Garfield, *et al.* 1987. *Cognitive science: An introduction.* Cambridge, MA: MIT Press. (Cited in Dawson (1997, in press).)

Tager-Flusberg, Helen. 1993. What Language Reveals about the Understanding of Minds in Children with Autism. In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press. 138-157.

Thelen, Esther and L. Smith. 1994. *A Dynamic Systems Approach to the Development of Cognition and Action.* Cambridge, Mass.: MIT Press.

Tolstoy, Leo. 1960. *The Death of Ivan Ilych and Other Stories.* Louise and Aylmer Maude trans. London: Oxford University Press.

Torrance, Steve. 1999. Real-World Embedding and Traditional Artificial Intelligence. In *Perspectives on Cognitive Science: Theories, Experiments, and Foundations.* Eds. Janet Wiles and Terry Dartnall. Stamford, Connecticut: Ablex.

Twardowski, Kasimir. 1894/1977. *On the Content and Object of Presentations.* R. Grossman trans. The Hague: Martinus Nijhoff.

van Gelder, Tim. 1995. What Might Cognition be, if Not Computation? *Journal of Philosophy* 92(7): 345-381.

Varela, Francisco J. 1979. *Principles of Biological Autonomy.* New York: Elsevier (North Holland).

Varela, Francisco J., Evan Thompson and Eleanor Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge Mass.: MIT Press.

Vera, Alonso H. and Herbert Simon. 1993. Situated Action: A Symbolic Interpretation. *Cognitive Science* 17: 7-48.

Von Eckardt, Barbara. 1993. *What is Cognitive Science?* Cambridge, Massachusetts: MIT Press, A Bradford Book.

Von Uexkull, Jakob. 1934. A Stroll Through the Worlds of Animals and Men: A Picture Book of Invisible Worlds. In *Instinctive Behaviour* Ed. K. Lashley. International Universities Press.

Wellman, Henry M. 1993. Early Understanding of Mind: the Normal Case. In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press. 10-39.

Whitaker, Randall. 1996a. *Introduction: Addressing Essential Circularity without Going in Circles.* http://www.informatik.umu.se/~rwhit/ObsWebIntro.html (Downloaded Sept, 1997).

Whitaker, Randall. 1996b. *Introductory Tutorial on Autopoiesis and Enaction.* http://www.informatik.umu.se/~rwhit/Tutorial.html (Downloaded Sept, 1997).

Whitehead, Alfred North. 1925. *Science and the Modern World.* New York: Macmillan.

Whiten, Andrew. 1993. Evolving a Theory of Mind: The Nature of Non-verbal Mentalism in Other Primates. In *Understanding Other Minds: Perspectives from Autism* Eds. Simon Baron-Cohen, Helen Tager-Flusberg and Donald J. Cohen. Oxford: Oxford University Press.

Whiten, Andrew. 1996a. Imitation, Pretence, and Mindreading: Secondary Representation in Comparative Primatology and Developmental Psychology. In *Reaching into Thought: the Minds of the Great Apes* Eds. A. E. Russon, K. A. Bard and S. T. Parker. Cambridge: Cambridge University Press.

Whiten, Andrew. 1996b. When does Smart Behaviour Reading become Mindreading? In *Theories of Theories of Mind* Eds. P. Carruthers and P. K. Smith. Cambridge: Cambridge University Press. 277-292.

Whiten, Andrew. 1997. The Machiavellian Mindreader. In *Machiavellian Intelligence II: Extensions and Evaluations* Eds. Andrew Whiten and Richard W. Byrne. Cambridge: Cambridge University Press. 144-173.

Whiten, Andrew. 1998. Evolutionary and Developmental Origins of the Mindreading System. In *Piaget, Evolution and Development* Eds. J. Langer and M. Killen. Lawrence Erlbaum.

Whiten, Andrew and Richard Byrne. 1988. Tactical deception in primates. *Behavior and Brain Sciences* 11: 233-73.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of some Recent Evolutionary Thought.* Princeton, NJ: Princeton University Press.

Wilson, David Sloan. 1997a. Introduction: Multilevel Selection Theory Comes of Age. *The American Naturalist* 150, Supplement(Supplement): S1-4.

Wilson, David Sloan and Elliot Sober. 1994. Re-introducing Group Selection to the Human Behavioural Sciences. *Behavioral and Brain Sciences* 17: 585-654.

Wilson, Edward Osborne. 1980. *Sociobiology.* Cambridge, Mass.: Belknap Press of Harvard University Press.

Wilson, Robert. 1997b. The Mind Beyond Itself. *Metarepresentation: the Tenth Annual Vancouver Cognitive Science Conference,* Simon Fraser University, Vancouver,

Winch, Peter. 1981. "Im Anfang war die Tat". In *Perspectives on Wittgenstein's Philosophy* Ed. Irving Block. Cambridge, Mass.: MIT Press. 157-178.

Winograd, Terry and Fernando Flores. 1986. *Understanding Computers and Cognition: A New Foundation for Design.* Norwood, New Jersey: Ablex Publishing Corporation.

Wittgenstein, Ludwig. 1958. *Philosophical Investigations.* Third edn. G. E. M. Anscombe trans. New York: Basil Blackwell and Mott. (Earlier edition published 1953, by Macmillan Publishing Co.)

Wittgenstein, Ludwig. 1958/1933-6. *The Blue and Brown Books*. Oxford: Basil Blackwell.

Wittgenstein, Ludwig. 1969. *On Certainty*. Denis Paul and G. E. M. Anscombe trans. Oxford: Blackwell.

Wittgenstein, Ludwig. 1980. *Culture and Value*. Peter Winch trans. Chicago: University of Chicago Press.

Wrathall, Mark and Sean Kelly. 1996. Existential Phenomenology and Cognitive Science. *The Electronic Journal of Analytic Philosophy* 4(spring): (EJAP is at: http://www.phil.indiana.edu/ejap/)

Wynne-Edwards, V. C. 1962. *Animal Dispersion in Relation to Social Behavior*. Edinburgh: Oliver and Boyd.

Zuboff, Arnold. 1996. The Story of a Brain. In *The Experience of Philosophy* Eds. Daniel Kolak and Raymond Martin. Wadsworth. 350-357.