Empiricism and the Epistemic Status of Imaging Technologies

by

Megan Catherine Delehanty

BSc. (Honors), University of Alberta, 1990

MSc., University of British Columbia, 1998

Submitted to the Graduate Faculty of

The University of Pittsburgh in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

Megan Catherine Delehanty

It was defended on

August 12, 2005

and approved by

James Bogen

Peter K. Machamer

John D. Norton

Simon Watkins

Sandra D. Mitchell
Dissertation Director

Empiricism and the Epistemic Status of Imaging Technologies

Megan C. Delehanty, PhD

University of Pittsburgh, 2005

This starting point for this project was the question of how to understand the epistemic status of mathematized imaging technologies such as positron emission tomography (PET) and confocal microscopy. These sorts of instruments play an increasingly important role in virtually all areas of biology and medicine. Some of these technologies have been widely celebrated as having revolutionized various fields of studies while others have been the target of substantial criticism. Thus, it is essential that we be able to assess these sorts of technologies as methods of producing evidence. They differ from one another in many respects, but one feature they all have in common is the use of multiple layers of statistical and mathematical processing that are essential to data production. This feature alone means that they do not fit neatly into any standard empiricist account of evidence. Yet this failure to be accommodated by philosophical accounts of good evidence does not indicate a general inadequacy on their part since, by many measures, they very often produce very high quality evidence. In order to understand how they can do so, we must look more closely at old philosophical questions concerning the role of experience and observation in acquiring knowledge about the external world. Doing so leads us to a new, grounded version of empiricism.

After distinguishing between a weaker and a stronger, anthropocentric version of empiricism, I argue that most contemporary accounts of observation are what I call *benchmark strategies* that, implicitly or explicitly, rely on the stronger version according to which human sense experience holds a place of unique privilege. They attempt to extend the bounds of observation

– and the epistemic privilege accorded to it – by establishing some type of relevant similarity to the benchmark of human perception. These accounts fail because they are unable to establish an epistemically motivated account of what relevant similarity consists of. The last best chance for any benchmark approach, and, indeed, for anthropocentric empiricism, is to supplement a benchmark strategy with a *grounding strategy*. Toward this end, I examine the Grounded Benchmark Criterion which defines relevant similarity to human perception in terms of the reliability-making features of human perception. This account, too, must fail due to our inability to specify these features given the current state of understanding of the human visual system. However, this failure reveals that it is reliability alone that is epistemically relevant, not any other sort of similarity to human perception.

Current accounts of reliability suffer from a number of difficulties, so I develop a novel account of reliability that is based on the concept of *granularity*. My account of reliability in terms of a *granularity match* both provides the means to refine the weaker version of empiricism and allows us to establish when and why imaging technologies are reliable. Finally, I use this account of granularity in examining the importance of the fact that the output of imaging technologies usually is images.

Acknowledgements

I would like to thank all of my wonderful committee members for their help and interest in this project. Special thanks are due to Sandy Mitchell for many, many conversations and for her support over the last 5 years, not only during the writing of this dissertation. Also, many thanks to Jim Bogen for reading every draft of every chapter and for the always productive (and fun) meetings we had about them. Many other people in the HPS department have also helped along the way – in particular, I'd like to thank Joann MacIntyre and Rita Levine for their always cheerful administrative help. Dan Steel and Wendy Parker also provided huge amounts of encouragement and mental health support along the way. And finally, I am enormously grateful to my parents for their love, encouragement, and undying support of all kinds as I made my long and winding way to this point. I am happy now.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

1.1. What we can learn from imaging technologies

Over the last 50 or 60 years there has been an enormous increase in the number and variety of visual imaging systems in the biological and medical sciences. The development of technologies such as electron microscopy (EM), X-ray computed tomography (CT), positron emission tomography (PET), X-ray crystallography, magnetic resonance imaging (MRI), atomic force microscopy, and many others has not only given scientists the ability to study objects at smaller spatiotemporal scales but also, in some cases, to investigate phenomena which had previously been inaccessible or incompletely accessible.[1] The development of these techniques, in general, can be interpreted as being motivated by the desire to see more or to see better – to increase magnification, spatial and/or temporal resolution, or to gain visual access to previously invisible (though not necessarily undetectable) phenomena. While there are, of course, enormous differences between various types of imaging systems, two features that are common to all of them[2] are, first, that the usual form in which the data is output is as more or less naturalistic images and, second, the need for computers that perform varying types and amounts of statistical and mathematical processing prior to the production of the image.

[1] By incompletely accessible, I mean that there were some properties of objects or phenomena which were (or came to be) recognized as crucial to biological explanations of specific phenomena but that could not be fully investigated with other, previously existing, tools. For instance, confocal laser scanning microscopy of living cells (CLSM) allows four-dimensional visualization of molecular scale processes about which information was previously accessible only by biochemical assay or by visualization of dead, fixed and stained cells (using EM or some other type of microscopy). Inaccessible phenomena include those that were spatially inaccessible (e.g. occurring within the brains of living humans) or those which we had no way to measure or observe, despite their spatiotemporal accessibility.

[2] Or, if not every recently developed imaging technology, at least of all those technologies with which I am concerned.

The fact that the images look essentially like pictures[3] of the objects they represent means that these technologies can seem to be providing us with a way of watching or seeing objects or events that are too small or are for other reasons inaccessible to human vision. As such, they can seem similar to photographs or video of everyday objects – forms of image that we often take to be good evidence – despite the fact that their means of production is usually very different. But facts about the means of production are absolutely central to whether or not an image is good evidence of some particular thing. Just as distance, the quality of the lighting, the type and quality of the film, and other such factors can affect how well or how poorly a photograph represents features of the object photographed, so too do various features of the production of images by these highly complex instruments affect whether, when, and of what they can potentially generate good evidence. One of the crucial features is the extensive mathematical and statistical processing that goes into the production of these images. Statistical treatment of data is required not only to interpret, but to *generate* the primary data. This, together with the fact that many of these instruments detect quantities such as radioactivity that are not directly detectable by any human sense, means that these instruments do not fit well with any empiricist account of evidence according to which distinctions between the observable and the unobservable and between theory and observation are essential to establishing the epistemic privilege associated with observation.

These imaging technologies and continuing improvements in them have been and continue to be central to many areas of research. As such, it is important to examine how they function as methods of producing evidence. The need for a philosophical examination of these instruments is particularly critical since there is considerable diversity of scientific opinion about the

[3] Though they look like quite low resolution pictures in some cases, such as positron emission technology (PET). Additionally, though the images are usually in color, the colors used often do not represent the color of the object but rather the value of the measured variable for that area.

evidential status of certain of these technologies. Some of these imaging systems are almost universally heralded as having revolutionized whole domains of study. Confocal microscopy in conjunction with the development of tags using naturally fluorescent proteins such as green fluorescent protein (GFP) is generally considered to have vastly increased the ability of cell biologists to ask a wide range of questions about events at the cellular and sub-cellular level. On the other side of the scale, significant doubts have been expressed about the value of evidence generated by positron emission tomography (PET) to identify areas of the brain that are involved in certain cognitive tasks. And, interestingly, somewhere in between these two is the use of PET for detection of different sorts of cancers. For some (e.g. non-small cell lung carcinoma), PET is generally accepted to be useful, for others (e.g. breast cancer) the quality of the evidence is debatable, while for yet others (e.g. bladder cancer)[4] PET cannot currently provide good evidence. If the problem were simply that these sorts of instruments sometimes fail to produce reliable data, then there would not be much of philosophical interest to be gained by examining them. However, the problem raised by these technologies is not that they sometimes fail – all instruments and our own senses are both fallible and useful only for certain applications – but that sometimes mathematized imaging technologies such as PET and confocal microscopy apparently *do* provide very good evidence despite the fact that neither fits any standard empiricist account of observation. By trying to assess how and when these technologies can provide good evidence about certain properties or features of objects, we are also forced to look more deeply into old philosophical problems concerning the role of sense experience and

[4] The most commonly used radiopharmaceutical in PET is fluorine-18 fluorodeoxyglucose (FDG). Detection of cancer using FDG is possible since most tumor cells have a higher rate of glycolysis than normal cells. Thus, their uptake of both unlabelled glucose and FDG is increased relative to healthy cells and they will show up as "hot" areas of increased FDG concentration. In the case of the bladder, however, even entirely normal tissue will show up as "hot" since, unlike glucose, FDG is not reabsorbed by the kidneys and is excreted into the urine, causing FDG concentration in the bladder to be high.

observation in getting knowledge about the natural world. This dissertation, then, will not only provide an account of when we can or cannot gain knowledge about the world using these sorts of technologies but will also develop a refined version of empiricism that identifies why the principles that empiricists have gotten right are right and rejects those aspects that are wrong.

1.2. Images as evidence

Scientific images and imaging technologies are a topic that has not received a great deal of philosophical attention in the past, though that situation has started to change over the last 15 years or so. There has recently been increasing attention paid to how various types of images – including graphs, diagrams, photographs, sketches, illustrations, and computer displays – play a role in science. This work, however, has focused on questions about what they represent or depict (Maienschein 1991), the relationship between diagrams and photographs or images and text (Rudwick 1976; Cambrosio, Jacobi, and Keating 1993; Krohn 1991; Lynch 1985, 1991; Myers 1988), what kind of role (if any) they play in arguments (Perini 2002, 2005; Kitcher and Varzi 2000) or in theories (Gilbert 1991; Taylor and Blum 1991; Taylor 1991; Giaquinto 1994), and whether visual representations are ever really necessary (Ruse 1991; Griesemer 1991; Wimsatt 1991). What has only begun to be examined, however, is how they function as evidence and, in particular, to the epistemic role played by the means of production of an image (Bogen 2001, 2002). This is an especially important issue when dealing with images that serve as primary data[5] rather than those that act more as illustrations of a concept or hypothesis such as Sewall Wright's path diagrams (Ruse 1990, Griesemer 1991). For images such as those

[5] Later I will argue that it is the numerical data that is more properly thought of as the primary data. However, since the conversion from numerical data to image occurs by a simple, conventional translation, the distinction does not make a difference as far as the means of production of the data is concerned.

produced by PET and confocal microscopy, the question of whether they are good evidence – and what they are evidence of – must occur prior to these other sorts of concerns.

Certain types of images have generally been taken to be very good evidence in that, under the right conditions, they reliably preserve some set of features of the objects they represent. In particular, photographic evidence is usually believed to be an accurate reflection of visually accessible features of the world. Of course, there may be questions about the epistemic credentials of photographs: whether that particular object was photographed in a way that misrepresents some of its features (or was manipulated after its production to achieve the same end), or how well the object photographed really represents a broader sample or class of objects. These sorts of worries, though, are usually relatively easy to resolve, especially when the photographs in question are of medium-sized objects easily visible to the naked eye. Photographs of objects too small or too far away to be seen using our unaided vision are usually granted the same status as those of medium-sized objects, primarily in virtue of the means of their production being the same. But while this is a legitimate inference insofar as the means of production refers only to the operation of the camera (or other recording device), in the case of small or far away objects we also need to include the instruments that make these objects available to the camera as part of the means of production of the photograph. The microscope, telescope, or other tool that makes such objects able to be photographed by the same detection devices that work for medium-sized objects also need to be shown to be capable of producing structure-preserving data. This is where the real work of establishing the reliability of various types of photographic evidence lies. The same sorts of considerations apply to images such as those produced by PET that are not strictly photographic, but look somewhat photograph-like and are at least roughly naturalistic.

For an image to be naturalistic or photograph-like is for it to preserve at least the spatial properties of the object(s) represented.[6] These features are preserved if there is a sufficiently accurate and precise 2-dimensional or 3-dimensional mapping between the spatial properties of the object and the spatial properties of the representation. As will be discussed in Chapter 4, what counts as sufficiently precise and accurate will purpose relative, so there is no sharp distinction to be made between what counts as photograph-like and what does not. A PET image is much more coarse-grained than many photographs of similar-sized objects – the spatial distribution of radioactivity in the data is represented much less precisely than the actual distribution in the object –but the relevant spatial features of the object are still spatially represented in the PET image, so it is included as photograph-like. On the other hand, a graph that shows the radiation intensity on the *y*-axis and an axis through the object (say, moving from top to bottom) on the *x*-axis will clearly not be photograph-like.

I take spatial features to refer primarily to size, shape, and relative position of objects. Other visually accessible features of objects may also be preserved in some cases, though they need not be. In particular, color (wavelength), and, in the case of the video, temporal relationships and motion may be represented and, so, potentially preserved, but these will not apply in all cases.[7] I will refer to this set of properties (spatial, temporal, color) of an object that are at least potentially represented in a photograph or other naturalistic image as *structural properties*. I intend this not in contrast with functional or secondary properties, but only as a way to distinguish the set of features of objects that are potentially visually accessible from those that are not, such as mass. Because objects that are not actually visually accessible to us – positrons, for instance – possess a subset of these properties, I do not want to refer to the set simply as

[6] From the point of view of the photographer or detection system.
[7] For instance, a black and white photograph does not represent color, though it may represent differences in tone or hue.

visually accessible properties since that would seem to indicate that we can straightforwardly see things such as the location of positrons. It is important to identify these properties since for imaging technologies to produce good or reliable evidence will turn out to be, in part, for them to generate data that is structure-preserving with respect to those properties that are represented by that particular instrument. A discussion of reliability must wait for later, but for now what matters is that there are some features of objects that we can get information[8] about by looking at the objects themselves or at photograph-like images of those objects. In the discussion of sense experience and observation that follows, I will be concerned primarily with visual perception rather than other human sensory modalities. Accordingly, it will be structural properties that are to be perceived or observed.

The division between what is observable – understood as accessible to sense experience – and what is unobservable serves as the dividing line between what the empiricist considers to be acceptable evidence and what is not. The motivation for drawing this distinction is to ensure epistemic security – to be sure that the data correctly represents the world (in whatever respects is important for the use to which we want to put the data). When we are concerned with visual observation, what we want, then, is that we get things right about visually accessible features of the world. Beginning with the idea that we need to interact epistemically with the world in order to get information about it and that the more direct and unmediated our evidence is, the less likely it is to be corrupted, empiricists have traditionally emphasized the importance of direct observation (unmediated sense experience) of the world. If instruments of various types (e.g. microscopes, telescopes) are allowed to count as unproblematic extensions of our senses and to produce good evidence, it is because they too generate structure-preserving data. The problem,

[8] Information here is used in a non-technical sense. Getting information about the structural properties of objects or groups of objects may be an end in itself, or it may serve another goal such as knowledge of the causal relationships that those objects stand in with respect to each other

then, is how to determine which instruments can do this. In some cases, it may be possible to compare the object viewed via the instrument to the object viewed via unaided vision, but in most cases of scientific instruments this is not the case. Thus, it is instead necessary to examine the means of production of the data in hopes of establishing that the process ensures preservation of structure.

1.3. Benchmark and grounding strategies

There are two ways to go about this analysis of the sorts of data-generation processes that can count as observation. The first is what I call a *benchmark strategy*. A benchmark strategy starts with the epistemic privilege that the empiricist accords to unaided human perception and attempts to extend the boundaries of observation by comparing the processes involved in the production of data by the instrument in question to those involved in unaided human perception. Instruments that use processes which are relevantly similar to the benchmark – human perception – are argued to share its epistemic status. The difficulty for this kind of strategy is in establishing what relevant similarity consists in. As we shall see later, different ideas for what counts as relevant similarity have been proposed. In order to make a principled choice between different versions of the benchmark strategy, it is necessary to switch to the second strategy: a *grounding strategy*. A grounding strategy attempts to identify the characteristics of human perception that make it a good source of evidence about certain kinds of things, to understand why it works (when it does), and to extend observation to include instruments that share those characteristics. By using a grounding strategy, it is possible to discover that the only relevant similarity to human perception that an instrument must share in order to produce good evidence is the fact that it produces reliable, structure-preserving data.[9] The question then becomes

[9] This is somewhat of a simplification; an account of reliability will be provided in Chapter 4.

whether extending the concept of observation to include all instruments that generate reliable data allows us to hold onto some form of empiricism.

Given that mathematized imaging technologies do not fit any current empiricist account and that there is ample reason to believe that there are cases where these technologies produce very good, reliable data, we can certainly redefine observation in a way that can include these instruments. But is this equivalent to abandoning empiricism? The answer depends on whether the way in which observation needs to be redefined retains core elements of empiricism or whether the required redefinition forces us to abandon central tenets that are constitutive of empiricism. To decide this, it is obviously necessary to briefly extract what the core principles of empiricism are.

1.4. Empiricism

It is difficult to identify empiricism as a single doctrine since the term has been used to refer to a number of considerably different approaches over the past several centuries. What all of them have in common, however, is an emphasis on the fundamental role of sense experience in acquiring beliefs and knowledge. A recent attempt to provide a definition of empiricism is as follows:

> "Empiricists believe that knowledge can only be obtained through the use of the senses to find out about the world and not by the use of pure thought or reason; in other words, the way to arrive at justified beliefs about the world is to obtain evidence by making observations and gathering data." (Ladyman 2002, 21)

While this is a good general account of the empiricist position, for our purpose here it will be necessary to look a bit more closely at the various forms empiricism has taken and, in particular, at what is motivating the insistence on sense experience. What is it that sense experience is

supposed to get us that reason or intuition cannot? And why or how does sense experience gain this advantage? Essentially this is to apply a grounding strategy to empiricism itself. Answering these questions by looking briefly at the history of empiricism will allow us to distinguish between two importantly different interpretations of the above description that are reflected in current debates both about observation and the epistemic position of mathematized imaging technologies. The first interpretation takes sense experience of some sort to be necessary since this is how we make epistemic contact with the external world.[10] We gather information by looking at things, by listening, by touching, and so on. These are forms of causal interaction with the world without which we could not discover its properties. This interpretation could be accepted by any empiricist since all it does is distinguish reason from sense experience and acknowledge that our senses happen to be the way humans have of epistemically interacting with the external world.[11] This interpretation essentially identifies a causal difference between sense experience and reason or intuition and finds that an epistemic difference attaches to that causal difference. It does not claim that the processes involved in human sensory capacities are somehow uniquely privileged ways of getting information about the world, but simply that they are to be preferred to the use of thought or reason alone and that they happen to be the proximal ways available to us to take in information from the world. The distal processes that contribute to our interaction with the world (e.g. the instruments that produce the image that we then look at) must be structure-preserving, but this doesn't require any similarity to unaided sense experience.

[10] This interpretation of can be found in current discussions of empiricism. For instance, John Norton introduces a discussion of why thought experiments are not problematic for empiricism as follows: "The essential element in experimentation is the natural world. We learn about the natural world by watching what it does in some contrived circumstance. Just imagining what the world might do if we were to manipulate it in this way or that would seem futile, since it omits this essential element." (2004, 44).

[11] I am concerned here with knowledge of the natural world. One could be an empiricist, rationalist, realist, etc. about different subjects such as morality or mathematics without committing oneself to holding the same position for physical or biological sciences.

This contrasts with the second interpretation which, as we shall see later**,** is reflected in many current accounts of observation such as those of van Fraassen (1980), Shapere (19820, and Hacking (1983).[12] This interpretation also identifies a difference in causal processes and a corresponding epistemic difference between sense experience and reason, but additionally identifies the processes involved in human sense experience as having a special epistemic privilege when compared not only to the use of reason alone, but to other processes that we might use to extend our perceptual capacities beyond the unaided use of our native sensory faculties. While the first interpretation drew a distinction between sense experience and reason without specifying the processes of interaction or involvement with the world that precede our direct sense experience of some object or event (i.e. it does not automatically distinguish between seeing the readout of some machine that measures the size of very small objects and visually comparing the size of medium-sized objects with your naked eye), the second interpretation implicitly makes a three-fold distinction between reason, processes of (or similar to) human sensory perception, and processes of interacting with the world that may end in sense experience but, because of their earlier unlikeness to human perception, do not share the epistemic privilege accorded to sense experience. In the chapters that follow, I will argue that the second, anthropocentric interpretation, though it is more prevalent, is unjustified and that empiricism can be upheld only if we use the first, weaker interpretation. The argument, in brief, is that there is no principled way to restrict epistemic privilege to those modes of interacting with the world that bear a certain kind of physical or causal similarity to human sensory modalities. By looking at what motivates empiricism, we can identify what sense experience is supposed to achieve: information about accessible features (those accessible via a particular sense modality

[12] This view is made explicit by van Fraassen and is implicit is some of the arguments made by Hacking and Shapere.

or instrument) of the world that is transmitted reliably via the processes involved in that modality. It is true that, when used under the proper conditions, sense experience usually achieves this, but it is not the case that this sort of reliability has as a necessary condition that the process of acquiring this information bear any particular sort of physical or causal similarity to unaided human senses.

In our examination of imaging technologies, we will see that without a more careful analysis of what is right about empiricism and why, we are unable to assess these instruments. Thus, an investigation of these instruments not only leads to a better understanding of the technologies but also forces us to develop a more refined version of empiricism. We will see that the second, anthropocentric version of empiricism cannot be maintained, but that the more refined, grounded empiricism can be.

If we look at the history of empiricism, we can identify two main versions: that of the British empiricists of the 17th and 18th centuries and that of 20th century empiricists including both logical positivists and contemporary empiricists such as van Fraassen.[13] Each shares the insistence on the role of sense experience noted in the earlier definition, but places different emphasis on two goals that are to be achieved by sense experience: 1) making contact with the external world, and 2) keeping out sources of error. For the British empiricists, the rejection of innate ideas was a key aspect of empiricism. As such, it contrasted primarily with rationalism and placed heavier emphasis on the first goal. Twentieth century empiricists, in contrast focuses more on the distinction between the observable and the unobservable, is contrasted more with realism than with rationalism, and emphasizes the second goal.

[13] This is not intended as a comprehensive account of the history of empiricism but only as a way to identify the central themes that have characterized empiricism and, in particular, to locate claims that indicate the first or second interpretation of empiricism as described above. There are obviously many differences between individuals whom I have grouped together – van Fraassen, for instance rejects the verification principle of the logical positivists – but the members of each of the two groups are united in which of two goals of empiricism is emphasized.

British Empiricists

Since I want here only to provide a general outline of empiricism as conceived of in different periods, I will illustrate the common themes of British empiricism by reference primarily to Locke with a lesser amount of attention to Hume. There are, of course, differences between Locke and Hume, but it will be more important here to focus on the commonalities in order to understand the motivation behind the position. The consensus position of the British empiricists was that if we have knowledge of a particular subject – and we may simply not have knowledge in some domains – then our knowledge is dependent on sense experience. Sense experience is our *only* source of primary or simple ideas – the innate concepts that rationalists took us to have were soundly rejected by both Locke and Hume. Though reason certainly plays a role in knowledge – in comparing ideas or combining simple ideas into complex ones, for instance – it is not by itself capable of giving us knowledge. It follows, then, that reason cannot give us knowledge with a higher degree of warrant, as rationalists claimed.

For Locke, ideas are of two kinds: simple and complex. We can get simple ideas only from experience. Prior to any experience, the mind is "white paper, void of all Characters, without any *Ideas*" (Essay II.i.2, p. 104). But then "*our Senses*, conversant about particular sensible Objects, do *convey into the Mind*, several distinct *Perceptions* of things … And thus we come by those *Ideas*, we have of *Yellow*, *White*, *Heat*, *Cold*, *Hard*, *Bitter*, *Sweet*" (Essay II.i.3, p. 104). All the content of our minds must ultimately be derived from experience: "All those sublime Thoughts, which towre above the Clouds, and reach as high as Heaven it self, take their Rise and Footing here" (Essay II.i.24, p. 118). This is not to say that we can have no idea of anything of which we have had no experience, however. Once we have gained from experience a number of simple ideas, our minds can derive from them complex ideas by the mental operations of "*Enlarging*, *Compounding*, and *Abstracting*" (Essay II.ii.22:117). Experience is of two kinds:

sensation and reflection. Sensation is what is crucial for getting knowledge of the external world; reflection is essentially an inner sense that makes us aware of the operations of our own mind. According to Locke, we can get some ideas only from reflection, but others come only from sensation.

But this is not to say that reason has no role in the production of knowledge. Knowledge, Locke says, is "*the perception of the connexion and agreement or disagreement and repugnancy of any of our Ideas*" (Essay IV.i.2, p. 525). It comes in three degrees: intuitive, demonstrative, and sensitive. When these perceptions of connections are very immediate and direct they can seem to be derived from reason or intuition alone; this is intuitive knowledge. When "the Mind perceives the Agreement or Disagreement of two *Ideas* immediately by themselves, without the intervention of any other" we can directly perceive "that *Three* are more than *Two*, and equal to *One* and *Two*" (Essay IV.ii.1, p. 530-31). Though such knowledge *seems* intuitive, it is nevertheless based on ideas that were first obtained by experience. In other cases, the connection between the two ideas is indirect and is mediated by other ideas (demonstrative knowledge) or is instead based on knowledge of the present existence of something in the world that corresponds to our current ideas (sensitive knowledge).

To these three degrees of knowledge are four types of proposition based on the type of connection between ideas. The sort that most concerns us here is general propositions about the properties of substances such as gold or silver. Locke takes it to be implausible that our knowledge of physical properties of gold (e.g. that it does not dissolve in nitric acid) and silver (that it does dissolve in nitric acid) is based on our perception of connections between our ideas. In such cases, it seems that reasoning about ideas doesn't play a role and that we are "left only to Observation and Experiment" (Essay IV.iii.28, p. 558). Accordingly, since knowledge requires

perception of connections between ideas, Locke declares that these are not cases of knowledge but of "belief" or "opinion" (Essay IV.xv.3:655). In these cases, we must use judgment about probabilities to determine what to believe. The central problem identified by Locke for our knowledge of the external world, or natural philosophy, is that there is a distinction between real and nominal essences of things. Our ideas of substances (such as gold and silver) are not of their real essences but only of their nominal essences while our ideas of triangles, for instance, are often, if not always, of their real essence. The details need not concern us here, but basically Locke insists that the sensible qualities of substances are not the features that make them fundamentally what they are. Only when our ideas are of real essences can we attain "certain and universal Knowledge" (Essay IV.iii.29:559) by the methods of intuition and demonstration. By this standard, "natural Philosophy is not capable of being made a Science" (Essay IV.xii.10, p. 645) since we cannot acquire real knowledge but are wholly dependent on beliefs formed on the basis of experience:

> "Substances afford Matter of very little general Knowledge; and the bare Contemplation of their abstract *Ideas*, will carry us but a very little way in the search of Truth and Certainty … *Experience here must teach me*, what reason cannot: and 'tis by trying alone, that I can certainly know, what other Qualities co-exist with those of my complex Idea, *v.g.* whether that *yellow, heavy, fusible* Body, I call *Gold*, be *malleable*, or no; which Experience … makes me not certain, that it is so, in all, or in any other *yellow, heavy, fusible* Bodies, but that which I have tried … Because the other Properties of such Bodies, depending not on these, but on that unknown real Essence, on which these also depend, we cannot by them discover the rest." (Essay IV.xii.9, p. 644)

Thus, we find in Locke the idea that the only method appropriate to the investigation of the external world, of the properties of substances, is observation. This he shares with both the logical positivists and contemporary empiricists such as van Fraassen, though for Locke the problem is with our understanding while for the logical positivists it is that the world is entirely

contingent so that observation and experiment are the only ways to investigate it. There are, of course, other differences – e.g. while the later logical positivists will insist that a priori knowledge of the sort that Locke refers to as "knowledge" has no real content, Locke firmly denies this, these differences can be set aside here..

In the case of the ideas we get from sensation, our knowledge can only be based on ideas formed on the basis of connections between apparent qualities of external objects rather than between ideas for which we create an internal ideal and those for which the comparison is ultimately to be made with some external object that we can only understand inadequately. This leads to the claim that our knowledge of the external world is inferior to our knowledge of other domains such as morality and mathematics. This is not to say, however, that reason or reflection provide knowledge of the external world: they do not, it is just that our knowledge in other domains has a higher degree of certainty while our knowledge of the external world is properly thought of as opinion and is probabilistic rather than certain. But without sensation, we could have neither certain nor probabilistic knowledge. Locke further distinguishes two categories of probabilistic knowledge. The first has to do with matters of fact – things that are available to observation and experience. The second concerns knowledge of things that are not available to the senses – things like atoms that are smaller than the lower limits of our sensible capacities or the medium-scale features of other planets that are too far away to be sensed by us.[14]

Hume's version of empiricism shares much in common with Locke's. The raw materials for our mental operations are "impressions" and "ideas" which are distinguished by their degree of force and liveliness (Treatise, I.i.1, p. 5). Impressions are livelier and ideas more feeble – when we actually see a color our experience is much more vivid than if we merely think of it. Ideas

[14] (IV xvi 12 p. 665-6)

are derived from impressions and are essentially weaker copies of them: "… all our simple ideas in their first appearance are deriv'd from simple impressions, which are correspondent to them, and which they exactly represent." (Treatise, I.i.1, p. 4). Thus, again, all of our concepts are ultimately derived from sense experience. Moreover, our impressions are more certain than our ideas: "These impressions are all strong and sensible. They admit not of ambiguity. They are not only placed in a full light themselves, but may throw light on their correspondent ideas, which lie in obscurity." (Enquiry, p. 62). Hume similarly classifies all true propositions as either matters of fact or relations of ideas and claims that reason (induction and deduction) cannot provide us with substantive knowledge of the natural world:

> "All the objects of human reason or inquiry may naturally be divided into two kinds, to wit, "Relations of Ideas," and "Matters of Fact." Of the first are the sciences of Geometry, Algebra, and Arithmetic, and, in short, every affirmation which is either intuitively or demonstratively certain. That the square of the hypotenuse is equal to the square of the two sides is a proposition which expresses a relation between these numbers. Propositions of this kind are discoverable by the mere operation of thought, without dependence on what is existent anywhere in the universe. Though there never were a circle or triangle in nature, the truths demonstrated by Euclid would forever retain their certainty and evidence. Matters of fact, which are the second objects of human reason, are not ascertained in the same manner, nor is evidence of their truth, however great, of a like nature with the foregoing. The contrary of every matter of fact is still possible, because it can never imply a contradiction and is conceived by the mind with the same facility and distinctness as if ever so conformable to reality." (*Inquiry Concerning Human Understanding*, IV, 1, p. 40)

Like Locke, Hume admits the role of reason in producing knowledge, but insists that sense experience is necessarily prior to reason. Both refer almost exclusively to features of the world that are detectable by the unaided senses so it is difficult to assess what Hume or Locke would have to say about modern imaging technologies. There is one passage, however, in which Hume

refers to microscopes and telescopes making it possible for us to have impressions which were not possible with the naked eye:

> "Put a spot of ink upon paper, fix your eyes upon that spot, and retire to such a distance that at last you lose sight of it; 'tis plain, that the moment before it vanish'd the image or impression was perfectly indivisible. 'Tis not for want of light that the minute parts of distant bodies convey any sensible impression; but because they are remov'd beyond that distance, at which their impressions were reduc'd to a minimum, and were incapable of any farther diminution. A microscope or telescope, which renders them visible, produces not any new rays of light, but only spreads those, which always flow'd from them; and by that means both give parts to impressions, which to the naked eye appear simple and uncompounded, and advances to a *minimum*, what was previously imperceptible." (Treatise 1.ii.1, p. 27-28)

This passage could be read as indicating some support for a benchmark strategy since microscopes and telescopes do allow us to get impressions and in "spreading" rays of light seem to be described as operating in a way similar to human perception. At a minimum, this passage indicates that Hume did not restrict sense experience to unaided sense experience.

Twentieth century empiricists

The logical positivists of the first half of the twentieth century shared the British empiricists emphasis on the role of sense experience and the crucial role played by observation in acquiring knowledge about the natural world. Dismissive of metaphysics and claiming that many traditional philosophical problems were meaningless, the logical positivists are probably most closely associated with the verification principle. A clear statement of this principle can be found in Hempel:

> "It is a basic principle of contemporary empiricism that a sentence makes a cognitively significant assertion, and thus can be said to be either true or false, if and only if either (1) it is analytic or contradictory—in which case it is said to have purely logical meaning or significance—or else (2) it is capable, at least

> potentially, of test by experiential evidence—in which case it is said to have empirical meaning or significance." (Hempel 1965, 101)

The logical empiricists held that both reason (logic) and empirical (sense) experience are sources of knowledge. Unlike Locke, at least, they do not deny claims made on the basis of sense experience the status of knowledge. Instead, they deny that logical or a priori truths have any substantive content, something that Locke would deny. Despite this difference, the underlying motivation seems to be similar: if we want to make claims (whether taken to be belief or knowledge) about the natural world, we need to use sense experience. The logical positivists took this idea and applied it to their account of scientific theories. The language of a theory contains two kinds of terms: observation terms and theoretical terms. Observation terms refer to objects or properties that can be directly observed or measured (i.e. they are verifiable) while theoretical terms refer to objects or properties that we cannot observe or measure but can only infer from direct observations. Once again, direct observation, though not defined, seems to refer to the use of unaided human senses though things like reading the position of needles on various sorts of dials or meters are accepted (Hempel 1965, 127). The point of allowing such things to count as direct observation, however, seems to be connected to the issue of public accessibility and interobserver agreement, rather than specifying types of instruments that count as allowing direct observations. It is not made clear, in particular, whether what is directly observed is only the position of the needle or whether what the position of the needle represents (e.g. the temperature in a gas or the level of radioactivity in a test tube) is also to count as having been directly observed. It seems more consistent with the use of their overall use of the term "directly observed" to interpret it as referring to the observation of the needle position alone, however, since this more easily fits with the apparent motivation for the verification principle –

the idea that sensory experience is more secure than inferences made on the basis of that experience.

The verification principle turned out to be extremely problematic and is rejected by more contemporary empiricists such as van Fraassen. As his constructive empiricism will be examined in Chapter 2, I will not go into any detail here. Briefly, however, van Fraassen clearly adopts an anthropocentric empiricism, claiming that what counts as observable is observable by humans: "It is these limitations to which the 'able' in "observable" refers—our limitations, qua human beings." (1980, 17). And once again, stronger emphasis here seems to be placed on the avoidance of error through the addition of potentially error-ridden intermediaries such as microscopes than the need for sense experience in order to make contact with the world.

1.5. So what is empiricism?

The two motivating principles that have historically been associated with empiricism, though emphasized more or less at different times are the need for sense experience to make contact with the world and the desire to eliminate potential sources of error. While the latter is often reflected in the specification that observation must be direct, involving unaided sensory experience, the passage from Hume showed that this was not an absolute requirement. That passage alone is not sufficient to identify whether Hume would advocate a benchmark or a grounding strategy to extend observation beyond the use of our unaided senses, but there is nothing in either of the two principles that would require retaining any special role for sense experience except insofar as it is required proximally in order for us to interact with the external world. As long as it can be established that instruments of any variety can be used without introducing error, then empiricism need not restrict observation to those instruments that bear some relevant similarity to human perception. It will turn out that, in order to account for the

production of reliable evidence produced by mathematized imaging technologies, we must reject any type of anthropocentric empiricism, but we can still accept a grounded empiricism based on the two above principles.

1.6. Outline of the dissertation

In Chapter 2 I will argue that existing accounts of the scope of observation are what I have introduced as benchmark strategies – they attempt to extend the boundaries of observation by reference to primarily causal similarity to human perception (HP). Moreover, they exclude the neural components of the visual system (the *endpoint problem*) in trying to determine what relevant similarity consists of and, as a result, the usually exclude even unaided visual perception as observation.[15] Even supposing that these accounts could overcome the endpoint problem, they still must fail since they do not provide sufficient justification for preferring one criterion of relevant similarity over another. I claim that this is because a benchmark strategy, if it is ever to succeed, must be supplemented with a *grounding strategy* that justifies the choice of a particular criterion of causal similarity to the benchmark, human perception (HP), in terms of the epistemic role played by the causal similarity. Chapter 2 will also present a detailed case study of PET that casts doubt on the idea that causal similarity to HP is required to get maximally reliable data since the application of mathematical and statistical processing algorithms that seem to have no correlate in HP increases rather than decreases the reliability of the data.

What the empiricist needs from an account of observation is to connect the reliability of HP (when it is reliable) with certain kinds of causal processes with respect to which other instruments may or may nor be similar. Chapter 3 identifies the *grounded benchmark strategy* (GBC) as the last best hope for the empiricist to maintain that the epistemic privilege normally

[15] van Fraassen is the exception here since his standard is perfect identity. His account, however, fails to provide sufficient reason for why perfect identity is to be preferred as the epistemically relevant criterion.

ascribed to human perception can also apply to methods of data production that bear no causal similarity to HP. The GBC asserts that we can perceive via an apparatus if and only if the apparatus is similar to human perception with respect to those features that make HP reliable. However, as this chapter will show, our understanding of the human visual system is not (yet) sufficient for us to specify its reliability-making features. The argument for the reliability of HP cannot, at least at present, proceed in terms of the reliability of the processes involved, but must instead be based on our long experience with it and our ability to manipulate conditions in order to test that what our eyes are telling us is reliable. Since we cannot identify the epistemically-relevant causal features of HP, we also cannot use any sort of benchmark strategy, even the GBC, to extend the realm of epistemically privileged observation. But this does not settle the matter in favor of the anthropocentric empiricist since what was identified in the ultimate failure of the GBC was the fact that it is *only* the epistemic criterion of reliability that matters. As long as the evidence produced by some instrument is sufficiently reliable (and we have some means of determining the reliability of that instrument), it can share the epistemic privilege given to HP.

The notion of reliability is a problematic one, however. After reviewing a number of different accounts of reliability within epistemology and philosophy of science, Chapter 4 will lay out a new, pragmatic[16] account of reliability that is relative to both the sort of discriminations that are needed for a specific purpose, and the sorts of properties or features of the world that an instrument (including the human visual system) can get at. This account identifies reliability as both preservation of the structure of the object and a match between the granularity of the world at which a particular question is directed and the granularity of the instrument. Both of these features are to be understood partially in terms of finite probabilities. The second part of the

[16] I am using "pragmatic" not in the way it is used by American pragmatists such as Dewey and James, but to refer to the sense of being based in practice (cf. Mitchell 1997, 2000).

chapter identifies strategies that are available for assessing the reliability of mathematized imaging technologies, once again focusing on PET. It contends that we have more tools for doing this than are normally recognized and, so, that the quality of PET evidence for applications including some functional brain imaging purposes can be established.

After focusing on the means of production of images that serve as evidence in Chapters 2-4, Chapter 5 turns to an examination of the significance of the fact that the data is presented in the form of images. While it is sometimes the case that initial evidence generated by a particular instrument *must* be an image[17] – for example, an X-ray produced using traditional X-ray film – this is not true for many modern imaging technologies. For many mathematized imaging technology, the primary data can be thought of as the numerical value or intensity associated with each pixel or voxel. This data can then be converted into an image by assigning a particular grey level or color to specific intensity ranges and displaying the data in a 2-D or 3-D array. Yet images are the most common form in which this data is displayed. Why? In particular, is there any epistemic advantage to specific data display formats?

The fact that images might be interpreted as being part of a relevant similarity to human perception and, in virtue of that, epistemically important, will be rejected in Chapter 3. The account of reliability presented in Chapter 4 applies equally to numerical data or images, so no advantage can be found there. After surveying a few of the historical, sociological, and rhetorical reasons why these data might be preferentially displayed as images, Chapter 5 identifies two potential epistemic roles for images: cognitive accessibility and, especially for video images, access to causal information. Images unquestionably provide an easier way for us to grasp certain features of the data, especially larger scale relationships between parts of the data. They normally do so at the cost of reducing the apparent granularity of the representation,

[17] This is not to say that the image cannot later be represented in a different form, quantitative or otherwise.

but since the full granularity of the instrument can be maintained over limited portions of the data, this need not restrict the ability of the user to make the discriminations required by the question of interest.

The second potential advantage, the ability to get causal information from video data, is more difficult to address both because the nature of causation is philosophically problematic and because scientific claims about the advantages of live cell imaging don't (and don't need to) distinguish between the advantages that are due specifically to differences in data display format (given the same data content) and those due to differences in the experimental set-ups that result in very significant differences in the content of the data represented in different formats. Without a doubt, live cell imaging allows much more data to be obtained, most notably by increasing the temporal granularity of the data, and its advantages are proclaimed with good reason. However, it will be shown that, if we are careful to eliminate these content differences, our ability to extract causal information from the data is not dependent on the format in which the data is displayed.[18] In identifying the minimal conditions required to get causal information from the data acquired using imaging technologies, the crucial role of background information is highlighted. While it may be the case that a non-Humean account of causation that would allow us to see causal relationships is plausible at the macro scale, at the micro scale we can neither see nor recognize the sorts of causal processes that are the equivalent of Anscombe's pushing, pulling, knocking over, etc. We see only larger scale spatiotemporal interactions that must be supplemented with background information about the kinds of influence that certain kinds of entities can exert and the conditions under which they can do so. Since the spatiotemporal information is also available from static images or numerical data, there is no advantage to video format.

[18] Except insofar as there are differences due to cognitive accessibility.

2. Observation and the benchmark strategy

2.1. Introduction

In emphasizing the role of sense experience, empiricism can be interpreted as making one or both of the following claims: 1) that observation (the use of our sensory capacities in *some* way) is necessary for getting evidence and knowledge about the world, and 2) that sense experience - and observation when taken to include only instruments that bear some relevant similarity to unaided sense experience - provides us with evidence that has a uniquely high degree of epistemic privilege. While the first of these is certainly true,[19] for the goal of understanding whether and when mathematized imaging technologies can produce good evidence, this interpretation, as it stands, is insufficient. It doesn't have the conceptual resources to allow us to distinguish between different sorts of technologies or different applications of those technologies since in every case we will eventually *look* at the images they produce. If this interpretation is to be useful for understanding contemporary science, therefore, it must be considerably elaborated. The account of reliability that will be developed in Chapter 4 is such an elaboration. Before that occurs, though, the second interpretation will be shown to fail.

For this stronger interpretation to hold, we must either define the scope of observation as co-extensive with any and all methods of getting the highest quality evidence, or establish that some particular subset of instruments and methods for collecting data have some common feature – other than the epistemic quality of the data gathered using them – that ensures that they, and no methods that do not share this feature, generate evidence with a uniquely high degree of

[19] This I take to be true as long as any new capacities that we might imagine humans to have or acquire that would allow them access to certain features of the world (some sort of ESP, for instance) would be counted as perceptual capacities.

epistemic warrant. The first strategy makes little sense for the defender of anthropocentric empiricism since it immediately abandons the idea that it is some connection to or similarity with human sensory capacities that is the source of epistemic privilege.[20] The second is the one that is normally used (van Fraassen 1980; Shapere 1982; Hacking 1983) and offers at least a greater chance of success. The difficulty with such a strategy, however, is in trying to identify the common feature. Existing accounts are what I call *benchmark strategies*: they allow or disallow certain instruments as modes of observation according to whether or not they bear a relevant similarity to the epistemic benchmark of human visual perception. The problem, then, becomes how to understand relevant similarity. No imaging technology or other instrument is identical to the human visual system. Instruments such as microscopes, telescopes, and positron emission tomography (PET) are like human perception in some respects but unlike it in others. A light microscope, like the human visual system, makes use of light within the (humanly) visible range of the spectrum. Unlike the human system that 'sees' primarily reflected light, however, the microscope involves the use of diffraction (Hacking 1983, 194-5). It is not immediately obvious whether it is the similarity or the difference that should matter if we want to know if the light microscope is enough like the human visual system to count as a mode of observation. Thus, we need a principled reason to prefer a particular standard for what counts as relevant similarity.

Since the goal is to use relevant similarity to define an epistemically privileged class of instruments, the criterion of relevance had better be one that ensures that these instruments produce especially good evidence. Notice that the empiricist who wants to defend the second interpretation needs this class of instruments – those that count as methods of observation in

[20] This principle could, in theory, be recovered, but then it seems to be a backwards sort of strategy at best. The empiricist could work back from the set of epistemically excellent methods and show that they all have some relevant similarity to human sensory capacities, but this seems unlikely to produce any principled account of what relevant similarity is and more than likely to produce some sort of *ad hoc* account.

virtue of some type of similarity to human perception – to possess a uniquely high degree of epistemic warrant. If there are methods that lack this similarity yet share the same epistemic status, then the empiricist cannot hold onto the claim that the evidence gained by our senses (both unaided and aided by the class of relevantly similar instruments) has any unique status. In the set of possibilities laid out in Table 2.1, the empiricist[21] requires that all of our instruments fall into either box 1 (observational + privileged) or box 4 (not observational + not privileged). If there is no principled account of observation that leaves boxes 2 and 3 empty, then all that is left of empiricism is the first claim that we need to use our senses in some way to get knowledge (of whatever quality) about the world. Though true, this position is so weak that it hardly seems to deserve to be called a philosophical doctrine.

1. Observational + Privileged	**2.** Not Observational + Privileged
3. Observational + Not Privileged	**4.** Not Observational + Not Privileged

Table 2.1 Possibilities for the status of imaging technologies.

Showing that box 2 is *not* empty will fall to Chapter 3. This chapter will be focused on an evaluation of existing accounts of observation and the challenges that technologies such as PET create for any account that implicitly or explicitly assumes that an instrument must bear some sort of similarity to human perception in order to produce good evidence. Van Fraassen (1980), Shapere (1982) and Hacking (1983) have all attempted, in different ways, to define the scope of observation. This chapter will evaluate these existing proposals and ultimately argue that none is adequate. Assessing both what these proposals are trying to capture about the nature of observation and the ways in which they fail will make clear what the desiderata of a better account of observation are. This will be facilitated by an examination of PET. This heavily

[21] Unless otherwise specified, for the remainder of this chapter, I will use "empiricist" to refer to the anthropocentric empiricist who wants to defend the second interpretation on empiricism.

mathematized imaging technology is used to "see" into the living body and produces images that are complex epistemic objects, a hybrid between a standard photographic image and the output of a mathematical model. The examination of PET will help to make clear what the limitations of the various existing accounts of observation are and will highlight certain features that must be accommodated in any account of observation.

This chapter is organized in the following way. Section 2 examines current accounts of observation and identifies both what features of observation they are trying to capture and how each fails to do so. Section 3 then provides an introduction to PET and the features of it that are particularly interesting and informative in trying to understand what counts as observation. Section 4 examines in more detail how the signal and signal detection system and mathematical aspects of PET challenge existing accounts of observation. Finally, section 5 identifies the general features that an improved account of observation must have, in particular, that a benchmark strategy alone cannot succeed but must be supplemented by a grounding strategy.

2.2. The Scope of Observation

Existing accounts of observation (van Fraassen 1980; Shapere 1982; Hacking 1983) are versions of the *benchmark strategy*, according to which a process qualifies as observation if it is relevantly similar to unaided human perception.[22] Each in its own way is trying to capture two key aspects of observation:

1) observation must preserve[23] spatial and other accessible features of the observed object (including color or temporal structure in cases where this is relevant), and

[22] van Fraassen makes this explicit; it is implicit in the accounts of Shapere and Hacking.

[23] What they are preserved in will vary according to the type of imaging system we are concerned with. In the case of an imaging technology that outputs a photograph or photograph-like image (e.g. an electron micrograph), we are concerned with that image. In the case of direct, unaided human perception, the question becomes more difficult since there is considerable debate over the nature of mental representation. However, for the purposes of this chapter it does not matter whether or not the mental representation of an observed object is picture-like.

2) this preservation of features is generated specifically by the processes involved in observation.

These two conditions are intended to ensure the epistemic security of observation. A benchmark strategy attempts to claim that both of these aspects are satisfied for certain types of instrument by starting from the fact that unaided human perception, under appropriate conditions of use, is usually very good at preserving certain features of the external world such as spatial relationships. Assuming that it is not accidental, this preservation must be the result of the causal processes involved in visual perception. If a process is reliable in one instance, it is potentially, though not necessarily, reliable in other instances.[24] Therefore, if the same or similar processes are used by some instrument, it is at least potentially capable of producing as good evidence as does human perception.

That observation should preserve certain features of the world (here, visually accessible ones) seems unproblematic, though of course there is more to say about the degree of accuracy and precision required as well as which features must be preserved using a given imaging modality. The second condition, however, requires further clarification. In particular, which sort of processes are allowed needs to be specified since, as the previous chapter noted and as will be examined in much more detail later in this chapter, one of the primary concerns with imaging technologies such as PET is the extensive statistical and mathematical processing that is involved in the production of the data.

In addition to this need for clarification of what observation is intended to achieve, however, I argue that there are three general problems that are found in these existing proposals. First is the *benchmark problem.* The choices of a particular standard for what counts as relevant

[24] No one would deny that processes may be less than maximally reliable under different conditions.

similarity to unaided human perception have not been well justified and fail to explain the privileged epistemic status of those technologies that possess the relevant characteristic. Van Fraassen, Shapere, and Hacking all differ in their choice of standard and although each offers some justification for his own choice of standard, even Hacking's - the most successful of the three - is not ultimately satisfactory. In order to have a principled benchmark strategy, what is ultimately needed is to identify as relevantly similar specifically those processes that contribute to the preservation of whatever features must be preserved for a certain instrument (including our unaided senses).

The second problem is the *perception-reliability problem*. Although they differ in what counts as relevant similarity to human perception, all three accounts presuppose that similarity to human perceptual mechanisms and reliability coincide.[25] They begin with the premise that similarity to human perception establishes the potential for some instrument to be reliable since similar processes should, under similar conditions, produce similar results. Regardless of whether or not this is actually true, it is insufficient to establish that other processes which are entirely dissimilar to human perception lack the potential to produce equivalently good, spatial or other structure-preserving data

Third, is the *endpoint problem*. Observation is in each case taken to be what happens in the interval between the object and the retina and ignores the fact that visual perception does not simply consist of the retina as a light detector but also involves at least some neural or cognitive mechanisms. These mechanisms both need to be considered as being important to establishing the relevant similarity of other instruments and must not be excluded as dissimilar by some

[25] Once again, Hacking's account comes closest to succeeding here since he acknowledges that instruments can fail to be reliable and discusses methods we can use to test whether or not the data generated is reliable or not. However, he does not mention the reliability conditions of unaided human perception nor does he consider whether instruments that involve physical processes that are not similar to human perception can still be similarly reliable.

chosen standard of relevant similarity if that standard is to be able to classify even unaided perception as observation.

While each account of observation shares the general features and, to a greater or lesser extent, the problems noted above, each is significantly different from the others so it will be most helpful to examine each in turn. Van Fraassen's is both the simplest and the most problematic, so let us turn to it first.

2.2.1. Van Fraassen

My focus to this point has been, and will remain, on observation, but van Fraassen is instead interested primarily in when we can claim that something is observable. He claims that "X is observable if there are circumstances which are such that, if X is present to us under those circumstances, then we observe it" (1980, 16) and is unconcerned with the question of whether or not we actually do observe some observable object using anything other than unaided human perception. As long as there are conditions under which we can directly see a particular object, it is observable. This has consequences for the debate between the scientific realist and the constructive empiricist. If an object is observable, then for the constructive empiricist to accept a scientific theory (as empirically adequate), what the theory says about this sort of object must be right (van Fraassen 1980, 18). This holds even in cases where the data to be explained by a theory is not attained by unmediated human perception. He allows, for instance, that the moons of Jupiter are observable since we could, with our unaided vision, see them if we were able to catch a ride on a spaceship out to Jupiter. An object that can only be detected with the aid of a microscope, in contrast, van Fraassen does not count as observable since there are no circumstances under which such objects can be observed using our unaided vision.

It may seem as though van Fraassen's question of what is observable is quite distinct from my question of what the scope of observation is, but in defining "observable" van Fraassen is

also making a claim about what counts as observation. For him, there is a clear connection between what is observable and what is knowable. Claims about observables carry more epistemic weight: we can have better, more secure knowledge about them. There is an obvious question about how we get this particularly secure sort of knowledge in cases such as that of the moons of Jupiter where no human has ever directly observed anything about the object so that observability is purely observability in principle, but setting this worry aside for the moment, let us look at how van Fraassen can be understood to be applying a type of benchmark strategy.

Van Fraassen's criterion of observability not only relies on a benchmark strategy, but on a very strict one that takes relevant similarity to human perception to consist only of identity with unaided human perception. The use of such a strict standard of similarity inherent in the unaided human perception criterion seems unmotivated and certainly conflicts with common usage of the term "observable" according to which such things as cells and other objects detected with a standard light microscope at even low power are said to be observable. Why, then, does van Fraassen choose this standard? Although he does not make it explicit, it seems that the reasoning must be that perceptual knowledge is uniquely well grounded and unproblematic. There is, then, a strong presumed connection between unaided human perception and reliability. The fact that the observability of an object is tied to its potential observation by unmediated human perception and can never be established by, for instance, validation or confirmation of detected features of objects by several mediated forms of observation (for instance, by using a standard light microscope, a fluorescence microscope, or an electron microscope) assumes that direct human perception has a particularly high degree of reliability. And, further, that it is a degree of reliability that either cannot be matched or that we cannot know is matched by any form of mediated vision. This is an especially clear form of the perception-reliability problem.

The existence of such a strong connection between unmediated human perception and reliability needs an argument but none is provided. Furthermore, if claims about observables-in-principle such as the moons of Jupiter are to have the same status as claims about observables-in-practice, it must be possible for at least some mediated forms of perception to provide the same degree of reliability as direct human perception. Looking through a telescope at celestial bodies is, for van Fraasen, to get information about observables. The evidence obtained in this way is no worse that the evidence we would get about these same objects if we were to see them up close, with our unaided vision. There is not supposed to be an epistemic difference between the two cases, but there is clearly a difference in the processes involved. Looking through a telescope is not perfectly identical to looking at something with the naked eye. Thus, the exact identity standard for relevant similarity cannot be adequate even for what van Fraassen wants it to do.

But is there a more adequate standard for similarity that could be recovered from his examples of things that are observable and unobservable? The place to look would seem to be the alterations of the way the world actually is that are allowed or disallowed in classifying objects as observable. It may be that the standard of direct human perception requires not perfect identity but rather possession of some set of properties of the actual human visual system. Could we, for instance, shrink ourselves down and have eyes that could detect energy with extremely short wavelengths and so observe much smaller objects than we actually can? There is no indication that van Fraassen thinks that this is the case. We can relocate ourselves in the universe in ways that are not currently possible but we are not allowed to imagine modifications of our actual visual system. Van Fraassen's strict standard really does appear to be a standard of identity, and so does not escape the objections noted above.

The failure of the van Fraassen's unaided human perception criterion, however, does not establish the insufficiency of the benchmark strategy overall. The promise of accounts that allow extensions of human perception seems greater, especially in light of the fact that critics of empiricism have long argued that the requirement that the observable be restricted to the observable by the unaided human visual system is misguided (Maxwell 1962; Hacking 1983). Both Hacking (1983) and Shapere (1982) have suggested alternative accounts of observation that are far more lenient than van Fraassen's.

2.2.2. Shapere

Shapere describes his position as a descendant of the empiricist tradition and aims to resuscitate its central claim that all our knowledge rests on observation from the criticisms noted earlier. In doing so, he acknowledges that unaided human perception plays a role in increasingly few cases of what many – scientist and philosopher alike – take to be observation. What matters for Shapere is not whether some object or phenomenon is observed by an unaided human observer, as van Fraassen demands, but only that it is *directly* observed. "Direct", for Shapere, does not require the absence of an intervening instrument between the object being observed and the human who will ultimately look at or use the data. An entity, X, is directly observed if: 1) information is or can be received by an appropriate receptor, and 2) the information is or can be transmitted *without interference* (this is the meaning of direct) to the receptor from x (Shapere 1982, 492). These requirements seem to be trying to specify the two features of observation identified earlier: that spatial features and relationships[26] of the observed object must be preserved, and that this preservation must be generated specifically by the causal processes involved in observation. Most of the work is done by the second requirement though the first is clearly necessary to ensure that anything even remotely resembling observation is possible.

[26] And others such as color and motion when relevant.

Non-interference is intended to ensure both that spatial features are preserved and that this preservation is non-accidental and directly causally related to the object's having these spatial features. These features of non-interference, then, are what are supposed to justify Shapere's particular choice of benchmark.

The non-interference criterion serves in large part to ensure that observation is non-inferential. By "non-inferential", Shapere specifically does not intend to refer to the logical meaning most people would understand the term to have. Instead, he uses it to refer to reasoning where we have no specific reason to doubt the leaps we are making and conclusions we arrive at (1982, 517). How is non-interference is supposed to guarantee that direct observation is non-inferential in this sense? A crucial part of Shapere's account is his extremely broad understanding of information. He includes not just interactions due to light and other forms of electromagnetic radiation, but those due to nuclear strong forces, weak forces, and gravity as forms of information that can potentially play a role in observation. The examples Shapere gives to illustrate the presence and absence of interference are the reception of electromagnetic radiation (photons) and neutrinos, respectively, from the core of the sun. Given existing knowledge about the physical characteristics of electromagnetic processes, we can expect a photon to travel less than one centimeter before interacting with some particle and being scattered, absorbed, or re-radiated. By the time it reaches the surface of the sun (apparently after 100,000 to 1,000,000 years), so many interactions will have occurred that the original high frequency, short wavelength gamma ray will now be low-frequency, long wavelength light (Shapere 1982, 491). The vast majority of electromagnetic radiation from the sun's core is received on earth in this altered form and so, Shapere concludes, provides us with only indirect or inferential knowledge of the core of the sun. In contrast, neutrinos exhibit very weak

interactions with other matter and have an extremely low probability of experiencing any interaction (interference) even over the great distance from the core of the sun to the earth. Thus, as long as we have detectors on earth that can capture these neutrinos, they will provide us with direct information about the core of the sun and the detectors can be said to allow us to directly observe the core.

This account contains a serious problem: the non-interference criterion is far too strict. Surely Shapere does not want to claim that any interference whatsoever makes an observation indirect? This would mean that almost all observation that involves electromagnetic radiation would not count as direct observation. Even the gold standard of unaided human perception would be cast into doubt since not all the visible light reflected by an object is received by the retina without further interference.[27] But if the criterion is not read as being absolute (or as nearly absolute as the case of neutrinos), then there needs to be some way of determining what extent or type of interference is acceptable. How this should be done is not obvious, though one possibility is to allow interference whose extent and characteristics can be predicted and corrected for on the basis of secure background knowledge. This would be consistent with other aspects of Shapere's account in which he is happy to allow that the use of background knowledge that there is little specific reason to doubt is epistemically unproblematic and does not prevent direct observation. So, for instance, assumptions about which neutrinos will actually be detected by a particular system (in the case he mentions, only the highest energy neutrinos) and their relative frequency does not create problems for the claim that the central core of the sun can be directly observed (Shapere 1983, 494). But in cases where there is interference, how are we to get this secure background knowledge? In the absence of direct observation there doesn't seem

[27] Not to mention that if unaided human perception cannot be direct observation, it would seem as though we could never even directly observe the data collected by neutrino detectors.

to be any way to know what sort of correction factors need to be applied. The causal basis for declaring something to be directly observable, therefore, is problematic in Shapere's account. The epistemology, on the other hand is very clear. Direct observation provides the foundation for knowledge so if evidence is obtained by direct observation it is epistemically privileged.

2.2.3. Hacking

Hacking's discussion of observation occurs as part of a larger project aimed at finding a philosophy of experiment that avoids both the difficulties with the observation sentences of the logical positivist program and the theory-laden observation of Kuhn (1970) and Feyerabend (1978). His discussion of observation is not primarily intended to define its scope, but rather to identify the conditions of production under which we should accept experimental data. He downplays the role that observation alone plays in science and even states that "Observation has precious little to do with that question" (Hacking 1983, 181) (the question of what makes an experiment and the results obtained from it convincing). Nevertheless, he also asks "How far could one push the concept of seeing?" (1983, 207), and provides guidelines for how we should answer this question, so it is clear that he does have an account of the limits or scope of observation (or "seeing").

Hacking's account is by far the most successful of the three and makes a real attempt at going beyond the benchmark strategy to incorporate a grounding strategy although, in the end, the particular grounding strategy that he uses doesn't succeed. Briefly, by attempting to ground the epistemic status of observation in the reliability of physical processes involving waves (electromagnetic or other types), Hacking is unable to accommodate neurochemical aspects of the visual system. As a result, his description proves to be insufficient even as an account of unaided human perception.

Hacking's account resembles Shapere's in some respects. As with Shapere, the criteria Hacking sets out for observation are intended to capture the idea that observation should preserve spatial relationships and that this can reliably be done if there is a direct causal relationship between the observed object and the eventual representation of it. Also, like Shapere and in contrast to van Fraassen, Hacking argues that we need not restrict the range of the observable to that which is observable with the naked eye. He allows that observation is not limited to systems that make use of the normal physics of human vision (reflected light in the visible spectrum) but that a system that makes use of "any property of *any kind of wave* at all" (1983, 198) can count as "seeing" as long as there is a direct interaction between the source of the wave, an object, and a sequence of *physical events* that conclude in the production of an image of the object that can be considered a good mapping (one in which the spatial relationships, 2-D or 3-D, in the structure of the object are reproduced in the image) of the interactions between the wave and the object (1983, 207-8). The idea that observation must consist entirely of a series of physical events serves the same role as Shapere's non-interference criterion, though it is significantly less strict. If there is an uninterrupted chain of physical events linking the object to our viewing of it, then it is at least possible that spatial relationships will be preserved.

But not all sorts of physical events will reliably do this: some physical events might introduce distortion or error. Shapere's non-interference criterion was a way to eliminate this possibility: Hacking instead uses the idea of a good mapping. We can judge whether or not the mapping is a good one since we don't just passively look at things in a microscope, we *interfere* (1983, 189). Just as unaided vision is an active process, so too is observing things under the microscope. Both unaided vision and seeing through a microscope can go wrong. Practice, alteration of conditions, and even change of instruments are crucial to establishing the epistemic

credentials of microscopic observation. In order to judge a map to be good, we need to know: 1) which features do not represent features of the object under study, and 2) which features are or may be artifactual or aberrant. Hacking's solution to these problems is to confirm the structural properties using a different technique. Unlike van Fraassen, this other technique need not be unaided human perception. He allows, for instance, that we can confirm the existence of a certain feature of platelets – the so-called "dense bodies"- which show up on transmission electron microscope (TEM) scans of these cells by using a certain fluorescent stain and viewing the cells using a fluorescence microscope. When the same grid of cells is observed using both techniques, the dense bodies show up in the same spatial location. Thus, they are taken to be real features of platelets, not merely artifacts of TEM (1983, 200-1). This example works because: 1) TEM and fluorescence microscopy involve both different cell preparation techniques and different optics (but can be used with approximately the same degree of magnification and resolution), and 2) in these cases it is true that *using* these microscopes does not involve theory (except in the sense that all observation, including unaided human perception, involves theory) since choices among algorithms, decisions about whether and how to correct the data, etc. are not performed.

Hacking's account is valuable in that it provides a basis for determining what good data is (a good mapping), acknowledges that particular methods of observation can sometimes fail to produce good evidence, and identifies some strategies that can be used to try to identify and/or correct for artifacts. However, the conditions that he identifies as needing to be present for "seeing" are not able to ground the usual reliability of unaided human perception, let alone extend the concept to the use of instruments such as microscopes. The primary problem is what I identified as the *endpoint problem*. Hacking's black boxing of all aspects of human perception

that occur after light hits the retina leads him to think that the feature of observation that is (usually) likely to produce good mappings is the involvement of a series of physical interactions between a wave of some sort, the object observed, and the receptor. This is Hacking's benchmark criterion and his attempt to provide justification for it (i.e. to implement some type of grounding strategy) is based on the fact that electromagnetic radiation outside the visible range and sound waves behave in similar ways in their interactions with matter (being reflected, refracted, etc.) and that these interactions are regular and so can produce reliable data. As far as it goes, this is fine. However, once you open the black box and go beyond the retina into neural mechanisms involved in vision, it is no longer possible to insist that the output of "seeing" be produced by a series of physical events involving interactions between objects, waves, and receptors. Neural mechanisms are certainly physical, but they involve chemical and electrical events rather than interactions with waves. Therefore, the use of any kind of wave will not be enough to establish relevant similarity to human perception. The idea that we can test whether a good mapping has been achieved by a particular instrument, however, is valuable since it makes explicit the idea that observation need not produce incorrigible evidence. Thus, while Hacking does not specifically address the fact that even unaided human vision is not always reliable, his discussion of the ways we might go about checking the reliability of our aided or unaided observations provides at least a partial solution to the perception-reliability problem.

2.3. Key features of PET with respect to observation

The three general problems with existing accounts of observation are brought clearly into focus when one considers PET. With respect to the benchmark problem, PET seems to resemble human perception in some ways (e.g. the data is usually presented as images that appear to give us a "view" inside the human body or brain) but clearly differs from it in others (e.g. our eyes

cannot detect radiation nor do they actually implement complex statistical and mathematical algorithms)[28], thus demonstrating the need for reasons to prefer one standard over another. Regarding the perception-reliability problem, the case of PET will show very clearly that mechanisms dissimilar to human perception may be needed to improve reliability and obtain better quality data. Finally, the importance of the endpoint problem is highlighted by the statistical and mathematical processing involved in PET. If observation consists of purely physical interactions involving EM radiation or other types of waves, then PET is clearly not observation. And, if the empiricist is right, this means that PET data is not as good evidence as is data produced by methods that do count as observation. But if the visual system is understood to be more than the retinas as light detectors, then the contribution that neural mechanisms make to the epistemic status of unaided human perception must be included. These mechanisms are not based on physical interactions with waves and are often at least described in terms of algorithms and computations, so there is clearly a need for a different account of observation to adequately describe evidence production even by unaided human perception. Such an account may or may not extend to include PET, but it must recognize the contribution of different sorts of processes to the reliability of the human visual system – and so more possibilities for relevant similarity – than did the accounts described in section 2.2.

PET is a non-invasive tool for imaging molecular processes occurring throughout the body. It generates what appear to be relatively naturalistic pictures of the brain or the body that are used for a variety of purposes from localization of specific cognitive functions in the brain to assessment of areas of reduced blood flow or tissue viability in the heart to identification of cancerous lesions anywhere in the body. PET was first developed in the early 1970's, shortly after computed tomography (CT) and at about the same time as magnetic resonance imaging

[28] Though neural networks may be modeled as if this is what is actually happening.

(MRI) (Phelps et al., 1975). The fundamental difference between PET and these other widely used imaging tools is that while CT and MRI are anatomical imaging techniques and can be used to visualize *structural* features, PET is a *functional* imaging technique and is used to identify regions of altered molecular activity. While CT and MRI are also heavily mathematized imaging technologies and so share some of the same features as PET, the fact that PET images function rather than structure raises particular epistemological challenges. Structural features can be seen with the naked eye once the body is cut open (at least structural features of the size that fall within the resolving power of CT or MRI) but sub-cellular molecular processes cannot. Thus, imaging of biological function allows us to "see" something that can be visualized by no other means.[29]

Functional images are a very complex type of visual evidence and have recently drawn the interest of some philosophers.[30] PET images (Figure 2.1) look somewhat like a picture of a particular anatomical region but just how (or even whether) it is that they can be claimed to

[29] While the debate over observability is often tied to the issue of realism, I will not be concerned with ontological issues here but only with epistemological ones. The claim that one can "see" brain activity using PET may seem to be of a very different nature than the claim that one can "see" tumors with PET since obviously tumors are visible by other means – including with the naked eye once the body has been cut open. But the difference is much more subtle than this since it is not tumors per se that are seen with PET but altered metabolic activity that is characteristic of malignant cells. And this type of metabolic activity, for instance the rate of glycolysis in a cell, is not something which can be "seen" in anything like the way a tumor can be seen. There is a difference between the two, but this difference is connected to details about the tracers used and their connection to the phenomena of interest (brain activity or malignancy) as well as to the limits and characteristics of both spatial and temporal resolution of PET. These issues will be discussed in more detail later.

[30] See, for instance, Bogen (2000, 2002), van Orden and Paap (1997), Stuffelbeam and Bechtel (1997). This interest has focused exclusively on functional imaging of the brain [functional magnetic resonance imaging (fMRI) as well as PET] and, in particular, on the use of functional imaging for research purposes (especially mapping brain functions). I will instead be focusing primarily on clinical applications of PET. While early clinical use of PET was in neurology and cardiology, today its predominant clinical use is in oncology. While there are many interesting issues that arise in the context of brain mapping that do not apply in the case of oncology (in particular, the validity of assumptions about localization of cognitive functions and the need to perform intersubject comparisons to generate a map of an average human brain), these issues are not directly related to the questions about PET that I am concerned with here. The actual imaging technology is the same for both research and clinical applications, however. Thus, restricting my focus to the use of PET in oncology will allow me to disentangle questions about the technology from questions about cognitive theory.

represent the functioning body is far from obvious. [31] Furthermore, it is not clear whether or not this type of representation is observational or perceptual in nature and whether PET shares the epistemic privilege normally associated with perception. PET images are very different from

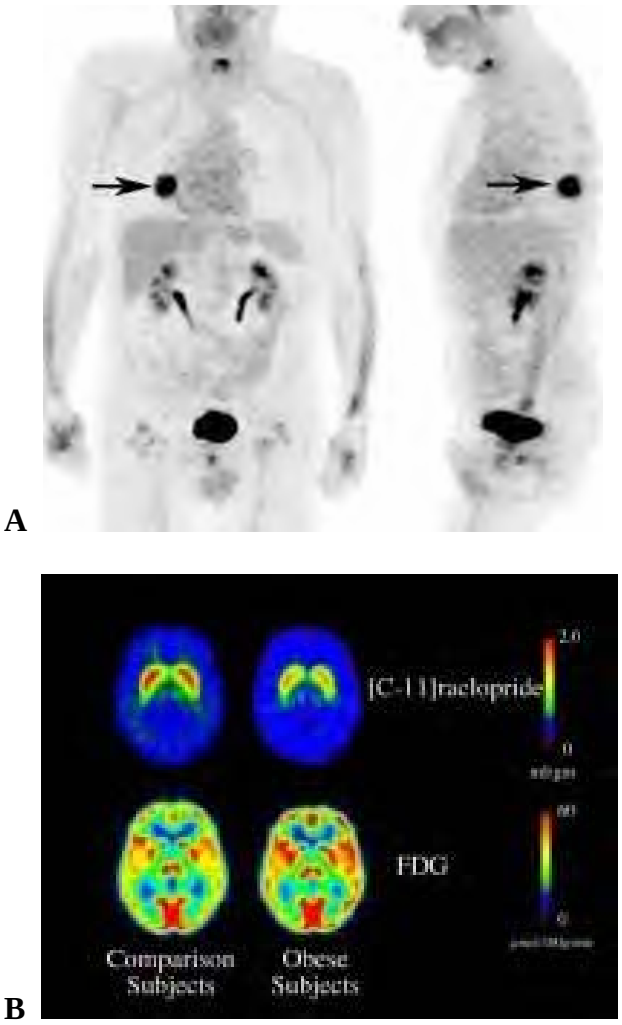


Figure 2.1 PET images.

PET images. Panel A is an FDG PET scan of an 80-year old man indicating with an arrow an area of hypermetabolic activity that was confirmed by biopsy to be non-small cell lung carcinoma (from Wang et al. 2001). Panel B shows a comparison of the number of dopamine receptors (raclopride is a dopamine D2 receptor antagonist) in normal and obese subjects (from Rohren, Turkington, and Coleman 2004).

[31] CT images also raise some epistemological questions since they, like PET, are dependent on reconstruction algorithms. Other aspects of CT, however, are less problematic than PET since they are, in essence, X-rays from which interference from planes other than the one of interest has been removed.

any sort of photographic image[32] since, first, they are neither produced by any kind of optical imaging technique, nor are the quantities they detect (511 keV photons) accessible to human senses. Thus, while they may appear to be straightforward photograph-like images, the physical process involved in their production is substantially different from that typically associated with perception. Second, and more importantly, they are produced only through the use of intensive mathematical and statistical processing. In virtue of this feature, they are not visual evidence in the relatively uncomplicated sense that a photograph is but are instead a hybrid epistemic object, sharing some features of a mathematical model or computer simulation in addition to some features of a standard (photographic) image. At the very least, this mathematical processing indicates that production of PET images includes some steps that are inferential, and depending on the nature of these inferences, it may involve interpretation rather than simply production of data. The former might be able to be incorporated onto a sophisticated empiricist viewpoint, but the latter is potentially more problematic. As a result of the nature of its signal detection and its extensive reliance on mathematical processing, PET is very difficult to accommodate in an empiricist view according to which observation and (direct) perceptual evidence are epistemically privileged ways of providing justification for our beliefs about the world. Since both aspects play an important role in putting up roadblocks to understanding the epistemic status of PET images according to existing accounts of observation, it will be necessary to examine both the detection system (what is detected and how) and the mathematical and statistical procedures that are involved in their production.

[32] I will use the term photographic image to refer not only to photographs but also to other images such as X-rays that are produced by direct optical means. The epistemic status of photographic images is not uncontroversial (even setting aside digital manipulation) and will be discussed in detail later, but for the present purpose the photographic image will be taken to be both epistemically much simpler than the PET image and, among durable, stable objects, the closest possible relative to perception.

While PET can be used to examine a variety of functional characteristics of different regions of the body, I will focus my discussion on its use in oncology. It will be important to examine three aspects in particular: the nature of the signal and its connection to the phenomena of interest, the signal detection system, and the process of image reconstruction from the raw data. In general terms, all of these features except the connection of the signal to the phenomenon of interest are common to all applications of PET. Thus, the present discussion of these aspects of PET in oncology will apply to other applications as well. However, because data correction and image reconstruction are complex mathematical and statistical tasks and there are many available alternatives in terms of strategy and algorithm, there is no single PET protocol or technique. The choice of strategy and of algorithm matters, and will be discussed in more detail in section 5. For now, however, only a general outline of the process will be given.

2.3.1. The signal and the phenomena

PET uses molecular tracers or probes that are specific to biochemical pathways or molecular targets in order to perform *in vivo* assays of physiological function. These probes are labeled with a positron-emitting isotope, are injected intravenously into the subject, and get distributed in the body according to the delivery, uptake, metabolism, and excretion characteristics of the probe. Depending on the task, a different probe as well as a different radioisotope will be required. Thus, for instance, brain imaging makes extensive use of ^{15}O while the use of PET in clinical oncology uses primarily an analog of glucose, ^{18}F-fluorodeoxyglucose (FDG)[33].

The most important characteristic of the radioisotope – other than its being a positron-emitter – is its half-life (Table 2.2). The half-life has a significant impact on the ability to detect particular sorts of phenomena. Too long a half-life means that detection of a sufficient number

[33] The development of new tracers is an important area of research since the lack of a suitable molecular probe makes PET studies of a particular phenomena inconclusive or impossible.

of counts[34] must either involve long scanning times (tying up the scanner as well as being inconvenient for the patient) or injection of larger amounts of radioactivity into the patient (which is clearly undesirable from a safety perspective). Too short a half-life produces other problems since it means that the isotope and labeled tracer must both be produced at the site where the PET scan is to occur. This requires a cyclotron (to generate the positron-emitting isotope) at the PET facility as well as the staff to operate the cyclotron and synthesize the tracer. In a clinical setting, and especially as the number of PET scanners in use grows and they are increasingly found away from major medical centers with associated research programs, these things are not always available. ^{18}F is by far the most commonly used in clinical applications since its longer half life allows it to be shipped from a regional manufacturing facility to hospitals or imaging centers.

Radioisotope	Half-life
Fluorine 18 (^{18}F)	110 minutes
Carbon 11 (^{11}C)	20 minutes
Nitrogen 13 (^{13}N)	10 minutes
Oxygen 15 (^{15}O)	122 seconds
Rubidium 82	75 seconds

Table 2.2 Radioisotopes used in PET.

While the characteristics of the radioisotope are important, even more important is the radiopharmaceutical probe or tracer into which it is incorporated. The ideal tracer for PET imaging in oncology is one that is specific for malignant cells (i.e. is not taken up by normal cells or by other disease processes) and provides a high contrast to background ratio (i.e it has a high uptake by tumor cells and no or negligible uptake by surrounding normal tissue). The kinetics of tracer uptake into specific areas is not crucial to oncology since tumors are stable entities at least when considered on the time scale of a PET scan, though kinetic features are very important to

[34] Since the longer the half-life, the less frequently an atom of the radioisotope decays.

applications where temporal resolution of molecular events is an important goal. This is most obviously the case in the use of PET for brain function mapping. While this issue is less acute for PET than for functional magnetic resonance imaging (fMRI),[35] there are real difficulties in trying to detect a fast event with a tracer that moves slower than that event. Because temporal resolution is not a concern for oncology, however, this point will not be elaborated on here though it will arise again in chapter 5.

As noted above, the radiopharmaceutical that is most commonly used in clinical PET imaging is an analog of glucose, ^{18}F-fluorodeoxyglucose (FDG). This compound was first described in the late 1970's (Gallagher et al., 1975) and its usefulness for PET studies is based on a particular metabolic feature of tumor cells that was first observed by the chemist Otto Warburg in 1930.[36] Warburg discovered that malignant cells have a much increased rate of glycolysis[37] relative to normal cells. Oxidative phosphorylation is almost entirely absent in tumor cells and anaerobic glycolysis is increased to make up for the loss of ATP from the oxidative pathway. This means that glucose is transported - via normal glucose transport proteins - into tumor cells at a higher rate than into normal cells. Importantly, glucose and FDG behave similarly in the initial stages of glucose metabolism and are distributed in tissue in proportion to glucose metabolic activity. FDG as well as glucose is recognized by the glucose

[35] The 1-2 second resolution of fMRI comes much closer to the millisecond time scale of neural events that the ~1 minute temporal resolution of PET.
[36] Warburg (1930).
[37] Glycolysis refers to the metabolic process by which glucose is broken down and converted to pyruvate (aerobic) or lactate (anaerobic). Increased glycolysis is not specific for malignancy, however. Increased glycolysis and glucose uptake also occurs with benign conditions such as infection, inflammation, and granulomatous disease. There are strategies to try to circumvent these limitations of FDG-PET and a significant amount of research is going into the development of radiopharmaceuticals that can make use of other features such as increased protein and DNA synthesis or elevated choline and phosphocholine levels in cell membranes that may be more specific to tumor cells.

transporter proteins and so is transported into tumor cells at a higher rate.[38] And once inside the cell, both glucose and FDG are phosphorylated by the enzyme hexokinase. At this point, however, glucose-6-phosphate continues through the glycolytic pathway whereas FDG-6-phosphate is effectively trapped inside the cell since tumor cells do not contain significant amounts of glucose-6-phosphatase, the enzyme that would dephosphorylate it.[39] Thus, FDG-6-phosphate accumulates inside cells in proportion to their rate of glycolysis and its distribution in normal and malignant cells can be imaged with PET. A semi-quantitative measure of glucose metabolism, the standardized uptake value (SUV), is often used to characterize lesions. The SUV is determined by defining an area of interest over the lesion and dividing the value (in microcuries per cubic centimeter) by the injected dose (in microcuries) divided by the patient's weight (in grams). The SUV associated with malignancy must be defined for each different type and location of tumor cells since the rate of glycolysis varies with different sorts of tumors as well as with different normal tissues (e.g. it is higher in muscle, the brain, and in the bladder).

2.3.2. Signal detection

As the radioactive atoms attached to the probe decay, they emit a positron and a neutrino. The neutrino passes out of the body without interacting and cannot be detected, but the positron rapidly loses energy in collisions with electrons in the tissue and within a short distance (usually less than 2 mm) annihilates with one of these electrons. The annihilation event produces two photons with an energy of 511 keV that are emitted with an angular separation of 180^{o}. A PET scanner is composed of 360^{o} 2-dimensional or 3-dimersional arrays of scintillation detectors that register "true" or coincidence events in which two photon interactions occur almost

[38] Though there are exceptions. FDG may have a low rate of uptake in some kinds of cancer such as low-grade lymphomas (Barrington and O'Doherty 2003). As well, certain normal tissues are known to have high rates of glucose and FDG uptake (e.g. the bowel, urinary tract, muscle, salivary glands, and lymphoid tissue).

[39] The main exception to this rule is the liver. Liver cells contain high concentrations of phosphatase enzymes so FDG-6-phosphate is dephosphorylated and cleared from the liver. This means that PET using FDG is not generally useful for liver cancers.

simultaneously on opposite sides of the head (or other region of the body). Scintillation counters work by coupling a dense crystalline material known as a scintillator with a photomultiplier tube.[40] When a gamma ray hits the scintillating crystal, some of the energy deposited is converted into a flash of visible light. The amount of light produced is low so in order to produce a measurable signal, a photomultiplier tube, optically coupled to one face of the scintillator is used to detect the light and amplify the signal (Knoll 1989). If the locations of both photons can be accurately detected, the line along which the annihilation (and, therefore, the positron emission) took place can be determined. These lines correspond to projections of the concentration of positron-labeled molecules in the body. By combining projections from many angles, the data can be reconstructed into cross-sectional images using reconstruction algorithms as will be described in the next section. The count density in the resulting images is then taken to represent the concentration of the positron-emitting probe in the tissue

However, before image reconstruction can begin, there are several sources of error - especially loss of true events - that must be corrected for at this point. These require the application of specific mathematical data corrections. These include correction for photon attenuation by tissue, correction for accidental or random coincidence events, differences in individual detector efficiency, and correction for detector dead time (processing of a detected event takes a finite amount of time and while this occurs the detector cannot detect another event). After correction factors are applied, the pixel intensity may still be in error due to resolution effects, thus partial volume correction must also be applied. At this point, image

[40] Different scintillating crystals can be used here and have significantly different characteristics. The exact nature of some data corrections depends on the type of scintillating crystal used as well as the way in which detectors are arranged (2-D or 3-D). These details, however, are less important than other mathematical aspects of PET and will not be discussed here.

reconstruction from the PET projection data can be performed. Attenuation correction is particularly important and will be discussed in more depth later in this chapter.

2.3.3. Image reconstruction

The task of image reconstruction arises because of the nature of the detection system described above. The result of using this form of detection system (and of detecting positrons to begin with) is that raw PET data, unlike an autoradiograph or an X-ray, is not an image of the object that was scanned: it is a set of numbers. The data is stored as a 2-D matrix known as a sinogram. The vertical axis represents the angle of the line of response and the horizontal axis represents the displacement from the center of the field of view (Figure 2.2). Each element in the sinogram represents the number of counts detected by a particular detector pair (members of a pair are located at 180^o from each other). Thus, any point in the (r, ϕ) coordinate system of the sinogram represents the count density at a particular point in the (x,y) plane being scanned. The sinogram is 2D but the body is 3D so somehow the spatial information needs to be recovered from the data. To do this – and to produce an image of the body from the sinogram - requires a reconstruction algorithm.

The basic problem is how to reconstruct a two- or three- dimensional image of the interior and exterior aspects of an object from data that consists of a large number of projections through the object. The mathematical techniques that were developed to accomplish this task are known as reconstruction algorithms. The primary problem is the superimposition of multiple planes in each data point (point on the photographic film or, later, the numerical figure representing a projection in a computer) resulting in blurring of the structure of interest by structures on front and behind it. Resolution of this problem thus required a strategy to eliminate the effect of radiodensity variations in planes other than the one of interest (or the one for which data is explicitly being sought). While I want to minimize as much as possible any mathematical

formalism, it is necessary to give a brief summary here in order for it to be clear what reconstruction algorithms need to do. As described above, the process of data collection must involve both translational and rotational scanning. Each plane of the object which is scanned, then, can be represented by an (x,y) coordinate system (Figure 1). The contribution of each point to the total detected signal is represented by the density function $f(x,y)$. 'Density' is not used in a technical sense here, but is used as a general term to cover both emission (PET) and transmission (CT) scanning. For CT, $f(x,y)$ actually represents the linear attenuation[41] coefficient μ; for PET, $f(x,y)$ is proportional to the radioisotope density. Ray-paths are described by an (r,s) coordinate system which is rotated by the same angle as the ray (see Fig. 2.2). Thus, each ray is specified by coordinates (r, ϕ), where ϕ is the angle of the ray with respect to the y-axis and r is its distance from the origin. The coordinate s represents path length along the ray. The integral of $f(x,y)$ along a ray (r, ϕ) is called the ray-projection p:

$$p(r,\phi) = \int_{r,\phi} f(x,y)ds \qquad (1)$$

For CT, p is proportional to the logarithm of the detector signal (because of physical considerations that are not important here). For PET, p is directly proportional to the total detector signal[42]. A complete set of ray-projections at a given angle is called a projection.

[41] The linear attenuation coefficient reflects the degree to which different materials (in this case, different tissues) attenuate, or block, the transmission of X-rays.

[42] Because MRI (and fMRI) is based on a different set of physical properties, the details of the detection system and characterization of the ray–paths differ. In particular, the data are collected in the Fourier domain (k-space) and include information on the amplitude, frequency, and phase of the precessing nuclei. Reconstruction, therefore, must involve fast Fourier transforms. However, many reconstruction algorithms also use Fourier transforms and are readily adapted for MRI and fMRI. The Fourier transform of a function $f(t)$ is proportional to: $\int_0^x f(t)\cos utdt$ or the corresponding function with sin ut. The transform allows a single non-periodic function to be expressed as a sum of trigonometrical functions of vanishingly small amplitudes. The transform itself pre-dates any work on reconstruction algorithms, having been developed by the French mathematician Jean Baptiste Joseph Fourier (1768-1830) (Glenn and Littler 1984, 74).

Ideally $f(x,y)$ is a continuous function and the number of projections is infinite. In practice, of course, $f(x,y)$ is calculated at a finite number of points from a finite number of projections. This is the starting point for the image reconstruction algorithms. These algorithms can be roughly divided into three classes based on the type of strategy they use to solve Equation (1). They are: 1) summation or back-projection, 2) iterative reconstruction, and 3) analytic reconstruction. There are interesting and important differences between the classes of algorithm, particularly with respect to the spatial resolution that they can provide, the artifacts or noise they

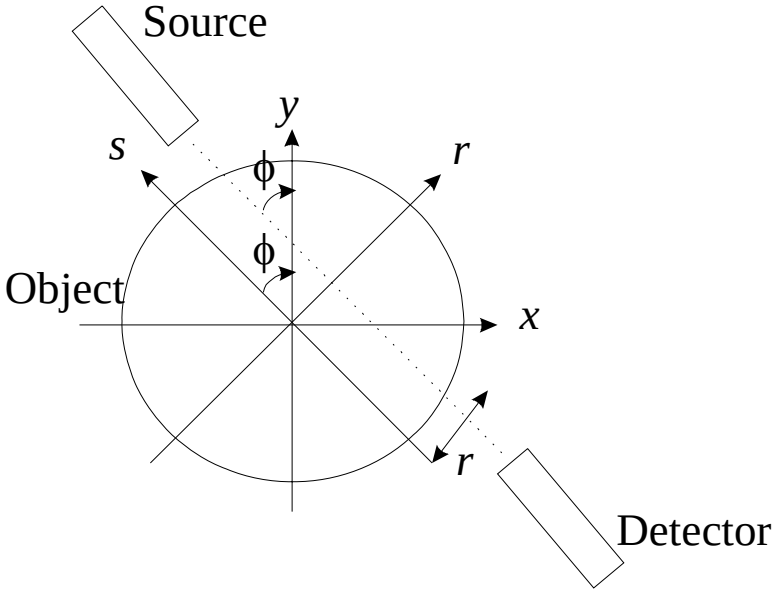


Figure 2.2 Coordinate system for describing ray paths and projections.
(Adapted from Brooks and DiChiro 1976).

introduce into the data, the assumptions they either require or are incompatible with, and the computational power they need to be useable in practice, I will not discuss them here.[43] For the present purpose, it is that a mathematical solution was required for this problem and that the solution is complex and involves trade-offs between many approximations and assumptions that matters.

2.4. Mathematical Aspects of PET

As is evident from the previous section, there are many layers of mathematical processing associated with PET and they clearly must be understood not only to contribute but to contribute very substantially to the process of PET image production.[44] Such procedures at the very least seem to be inferential. The problem with inference is that it can seem to blur the line between producing data and interpreting it and part of what has traditionally been taken to contribute to the privileged epistemic status of observation is that it does not involve interpretation.[45]

The crucial aspects of this extensive mathematical dependence is whether it automatically forces us to conclude that PET is more epistemically questionable than methods of data production that do not involve inference or whether only particular types of inference render PET data epistemically problematic. If some kinds of mathematical processing can be shown to

[43] An excellent review of these differences is provided by Brooks and DiChiro (1976). Gordon, Herman, and Johnson (1975) is a good, non-technical summary.

[44] It may seem that PET could unproblematically be considered observation on van Fraassen's account. To understand why PET cannot be used to observe tumors in the same way that a telescope allows us to observe the moons of Jupiter, recall that what PET is detecting is not actually tumors per se but (in the case of FDG-PET) increased glycolytic activity in cells. If what was detected were instead structural anatomical features of tumors of a sufficient size that the naked eye could discern them once the body was opened up and the tumor looked at in situ or removed and examined, then van Fraassen might be happy to concede that tumors are observable using PET. However, the situation here is closer to that of seeing parts of cells using a microscope. There is no context in which we could, with our unaided vision, observe either the build-up of FDG-6-phosphate inside tumor cells or other features of altered tumor cell metabolism. Thus, van Fraassen along with the traditional empiricist would have to conclude that PET cannot be used to observe cancer.

[45] Thus, for instance, Hacking's emphasis on the fact that using a microscope does not require knowing or applying optical theory (1983,169-209).

improve the quality of the data, then it seems that there is good reason not to exclude such processes from the epistemic evaluation of particular technologies. Whether an account of observation that uses a benchmark strategy can accommodate this processing as part of a relevant similarity to human perception, is another question. Shapere argues that direct observation can include certain types of inference (including mathematical) while Hacking, as noted above, excludes all but physical processes from the production of what can count as a good map and therefore as observation. As described in the previous section, there are many layers of mathematical processing involved in PET. Each layer, each type of correction and reconstruction algorithm, as well as their relative virtues and flaws is the topic of extensive continuing research. This section will focus on strategies for calculating and correcting attenuation since this is an area of current debate and one where the role of mathematical processing in preserving structural relationships is especially evident.

Attenuation occurs when the photons are absorbed or scattered while passing through the body, thus leading to the loss of detection of true events. For whole body PET, attenuation is the greatest source of loss of true counts. Approximately 10% of the 511 keV photons are lost for every 1 cm of tissue traversed. For whole-body PET, this means that the uptake of radioactive tracer is measured at only 5%-20% of the actual value if attenuation corrections are not applied (the loss is more severe the larger the patient). This can lead to very low, even negative tracer concentrations being assigned to some areas. For instance, consider a sphere of uniform radioisotope density and uniform attenuation. Without attenuation correction, the center of the sphere will appear to be of much lower radioisotope density than the outside.[46] In other words, the data will be consistent with a sphere of non-uniform radioisotope distribution, with the central region having a lower radioisotope density than the outside. It is also consistent with

[46] A mathematical proof of this is given in Bai et al. (2003), but the intuitive example will suffice here.

more than one different pattern of non-uniform attenuation. If attenuation is not measured accurately and appropriate corrections applied, therefore, it will not be possible to distinguish between several possibilities. In addition to quantitative inaccuracies, attenuation also causes specific and well-known types of qualitative artifacts. These include a prominent body outline that resembles high skin uptake, distortions of high-uptake areas such as the bladder (due to more attenuation in some directions than in others), and apparent high tracer uptake in low-attenuation areas such as the lung.

Attenuation creates problems in objects of both uniform and non-uniform composition, but the effects are more severe and the true tracer distribution harder to predict in objects of nonuniform attenuation. The amount of attenuation varies between tissue types (Rohren et al. 2004, 308) and depends on both density and the chemical composition of the tissue. Thus, since most soft tissues (blood, liver, pancreas, etc.) are over 70% water[47] and overall have very similar densities, their attenuation coefficients are approximately equal. Lung tissue has a very similar chemical composition to other tissues, but has a lesser proportion of water and a far greater proportion of air (especially when the lungs are fully expanded) than most soft tissues and hence has a lower attenuation value than other soft tissues. Adipose tissue has a slightly lower attenuation value because it is composed of more lipids and less water. Bone has large chemical and density differences in comparison to soft tissue and has a much higher attenuation rate. In the case of PET imaging of the brain, attenuation can be treated as constant; however, in whole body imaging this is clearly not always the case. Imaging of some regions of the body, in particular the abdomen as it is composed of primarily soft tissue types, can be treated as having uniform attenuation. However, most other areas of the body will contain areas of with very

[47] The grey matter of the brain is 70.6% water, the white matter 84.3%, the heart 80.0%, and blood 93.0%. In contrast, bone is 12.2% water.

different attention values, for example, soft tissue, bone, and lung in the thorax. Thus, one cannot assume a standard attenuation rate, and to correct for the amount of attention in a particular mixed tissue body area, the actual amount of attenuation for each projection must be measured. The standard way of measuring attenuation factors for PET has been to use a transmission scan prior to the emission scan. This involves using an external rotating positron-emitting source (normally germanium 68 or cesium 137) and measuring the proportion of the known positron emissions that are detected for each plane: the higher the proportion, the less the total attenuation for that plane. In dual modality PET/CT scanners, attenuation factors can be calculated much more quickly from the CT scan rather than a transmission PET scan.[48] Since attenuation values are energy dependent, the correction factor calculated from a CT scan at a mean photon energy of 70keV needs to be adjusted to the PET energy of 511keV. The scale factor is determined by the chemical composition (electron density) of a particular tissue. In theory, scaling would take into account even small differences between or within tissue types (for instance, different bone types have different fractional amounts of water-like tissue and dense cortex) but in practice, attenuation values above and below certain thresholds are treated using the same scaling factor. Whichever way these attenuation factors have been calculated, they can then be applied to the reconstruction of the PET emission data collected to correct for variable amounts of attenuation in each plane.

Given the problems associated with failure to apply attenuation corrections, one might think that their use would be an indispensable part of PET. However, this is not the case. In the recent clinical literature, there has been a considerable amount of debate about whether or not one

[48] The speed of this measurement is a great advantage, though breathing artifacts are a serious concern with this method, especially when the CT scan is performed with the lungs maximally expanded (i.e. with the patient holding their breath) and the PET is acquired with the patient breathing normally. It has been reported that in 300 patients with known liver tumors, PET/CT localized the tumor to the lung in approximately 2% of the cases (Osman et al. 2003).

should even attempt attenuation correction in whole-body PET imaging (Bengel, Ziegler, and Avril 1997; Bleckman et al. 1999; Farquhar et al. 1999; Wahl et al. 1997; Wahl 1999; Bai et al. 2003). The advantages of image reconstruction without attenuation correction have been claimed to be avoidance of the noise amplification that is inherent in attenuation correction (this can come from two sources: the noise multiplicative effect of the attenuation correction and the noise inherent in the correction factors themselves)[49], reduction of the patient scanning time (though this does not apply to PET/CT), avoidance of potential artifacts which arise from patient motion occurring between the transmission and emission scan, and improvement of the contrast to noise ratios for lesions because of reductions in the local background. As noted earlier, the disadvantages of not performing attenuation correction are quantitative and qualitative inaccuracies in determination in the shape and location of lesions. Bai et al. (2003) have attempted to resolve this dispute by performing a systematic survey of PET tumor detection with and without attenuation correction using simulation studies, a phantom experiment, and a patient experiment. Their results showed that lack of attenuation correction in areas of uniform attenuation can enhance contrast due to lowering of the local background level. This can facilitate detection of lesions. However, in areas of nonuniform attenuation (their work focused on the thorax in particular) image reconstruction without attenuation correction can result in the total absence of contrast between areas of background and of increased tracer uptake, thus making some lesions undetectable. The tracer concentration at which this effect occurs depends

[49] This is the most significant problem epistemologically. There are several methods that can be used to estimate the attenuation coefficients from the transmission scan data (or from CT in the case of a PET/CT scanner) but all of them result in the propagation of noise and measurement errors from the transmission data into the reconstructed map of attenuation coefficients and from there to the attenuation corrected emission data and eventually into the final reconstructed image of positron emission activity. Correcting for this noise and error propagation in turn is made difficult by the fact that the noise in the data is not linear but instead signal-dependent. This means that non-linear methods for reconstruction are required and these are more difficult and less well worked out than linear methods. In addition, non-linear methods are particularly affected by the fact that the spatial resolution of PET is spatially variant and object dependent (Lewitt and Matej 2003, 1606).

on the size, location, and density of the tumor but, importantly, is independent of the method used for reconstruction and of the acquisition mode (2-D or 3-D). In other words, *only* by using attenuation correction can these lesions be detected.

The fact that both calculation of attenuation and correcting for it in the data set rely on mathematical processes and assumptions seems problematic for Hacking since these are not physical processes and yet have a strong influence on the generation of the image. It seems that Hacking's account then must categorize PET as non-observational for this reason despite the fact that it could count as observational on the basis of the signal and signal detection. The problem with this is that there is no real reason given for why only physical processes are allowed. It may well be the case that the epistemic status of inferential procedures is not as good as that of (some) physically based processes, but we need to have some grounds for making this claim. A plausible reason that might be given on Hacking's behalf is that observation must be non-inferential and these mathematical processes clearly involve inference. Such a claim might be reasonable given the significant amount of work that has claimed that the epistemic privilege of unaided human perception is connected to the fact that it is non-inferential. The general claim, though, assumes that the addition of inferential reasoning to causal processes must always result in a process that is less accurate than the causal processes alone. The fact that some mathematical processes – inferences – clearly have a positive effect on the quality of PET evidence is difficult to reconcile with the view that *all* inference is problematic. Without attenuation correction, the quantitative and qualitative relationships in the PET image are simply wrong in many respects. But Bai et al showed that, despite worries about the noise-amplificative effects of attenuation corrections, they do produce what Hacking would certainly characterize as a better (if perhaps not good) map.

Shapere's account does not automatically exclude inferences from as long as we have no specific reason to doubt them. The basis for this is that just as background knowledge about the nature of the signal and detection system contributes to determining what counts, epistemically, as observation, so should secure background beliefs be allowed to contribute elsewhere. For instance, observation of neutrinos allows us to observe the composition of the core of the sun only with a chain of complex calculations based on the age of the sun, nuclear reactions, and stellar evolution (Shapere 1982, 517). Epistemically, we need to be concerned only with how secure these inferences are, not with the fact that they are inferences. That something is an inference does not determine its epistemic status. The question for PET, then, is how secure the mathematical inferences are. This is a difficult question to answer since each algorithm and each correction factor undeniably involves assumptions and simplifications that are known to be false. In part, the question can be addressed within the domain of mathematics alone, but at least in the case of PET in oncology, it is possible to perform validation studies of various types of algorithm and correction. As shown above in the discussion of attenuation factors, this can help to establish which mathematical processes involve secure or insecure inferences.

Suppose that some PET image i, is used as evidence for the existence and location of a tumor t. If i is to be considered good evidence of t, then i should have the characteristics (spatial, quantitative) it does only if there actually is a metabolically active tumor of a particular size and shape at a particular location. Unlike the case of brain imaging discussed by Bogen (2002), this sort of counterfactual dependence can be tested for oncological applications of PET. This applies to both the signal detection system and the mathematical processing. In the case of living patients, the diagnosis or staging of a tumor can be confirmed by visual inspection (upon surgery) and biopsy (this is, in fact, the standard means of validating PET and of estimating its

specificity and sensitivity for detecting tumors of particular kinds)[50]. PET can also be tested with the use of what are referred to as "phantoms". Phantoms are objects which are constructed to have specific known physical characteristics (shape, size, attenuation coefficient, and radioisotope density). These objects can be placed in a PET scanner and the results of using particular types of data correction or reconstruction algorithm assessed. Since the characteristics of the phantom are known, the characteristics of the image produced can be compared to those of the phantom. In this way PET in oncology can be assessed through investigation of the counterfactual dependence of the image characteristics on the characteristics of the object scanned. This dependence cannot be determined completely accurately since the case of scanning real humans involves an additional source of error: motion. The phantom does not move at all while an actual human engages in overall bodily (though everything is done to minimize this) as well as internal motion at least some of which is not under voluntary control. Imaging of the thorax is particularly vulnerable to motion-related sources of error since there is considerable motion associated with breathing as well as with the beating of the heart. Even if a breathing, heart-beating phantom were constructed, it would not be possible to match all of its motion characteristics (timing of breathing and heartbeat, volume of lung expansion, etc) to those of a given patient since these are not known quantities. Such motion does create real problems, resulting, for instance, in some lung tumors being incorrectly localized to the liver. However, the conditions under which such problems arise can normally be specified and either measures taken to minimize them or particular caution used in assessing the images produced.

[50] A recent meta-analysis of the literature on the use of PET to distinguish benign and malignant nodules showed that PET was 97% sensitive and 78% specific for malignancy (Gould et al. 2001). A preliminary study on using SUV as a semi-quantitative measure of the likelihood of malignancy found that by using a cut-off value of 2.5 the specificity of PET for a benign lesion was 100% and the sensitivity for a malignant lesion was 89% (Patz et al. 1993), though for lesions of under 1.5 cm partial volume effects make assessment more difficult since an SUV of <2.5 may be due to the limits of PET spatial resolution rather than the nature of the nodule (Matthies et al. 2002).

Additional discussion of strategies that can be used to assess the epistemic status of PET, including its mathematical aspects, will occur in Chapter 4. For now, it is important to notice that statistical and mathematical processes can be shown to often increase the quality of PET data. Thus, either these sorts of processes must somehow be included as relevantly similar to some aspect of human perception (if the anthropocentric form of empiricism is to be retained), or it must be concluded that similarity to human perception is irrelevant to the epistemic status of an imaging technology (in which case we can retain only the interpretation of empiricism that claims that we need sense experience to make causal contact with the world). Chapter 3 will argue that we must do the latter, then Chapter 4 will add to this minimal version of empiricism the resources to assess the evidential status of various imaging technologies.

2.5. Conclusion

Increasingly biology and medicine rely on heavily mathematized imaging systems such as PET. It is, therefore, important to understand how to assess the evidential status of the data (usually images) that such technologies produce. An interpretation of empiricism that takes sense experience to be crucial only because we need it in order to make any sort of contact with the external world has, as it stands, nothing to say about the epistemic importance of different sorts of causal processes that occur earlier in the chain of events that lead to our eventual sensory experience. It does not, in other words, allow us differentiate the evidential status of images produced by very different means. As such, it cannot help us to assess why (or even whether) attenuation-corrected images are better than non-attenuation corrected ones or if confocal videos of GFP-labelled cells are likely to be better evidence of some sorts of phenomena than PET

images may be of how the brain performs certain cognitive tasks. The anthropocentric interpretation of empiricism was theoretically able to help with such questions, but PET is not well accommodated by any existing account of observation that is based on this interpretation. Moreover, an examination of PET highlights certain difficulties that afflict each account. These problems are of three general kinds. First is the benchmark problem: the failure of an account of observation to provide a principled reason to prefer what it takes to be relevant similarity to human perception over other proposed standards. Second, the perception-reliability problem presupposes that reliability and similarity to human perception must coincide. Extending the bounds of observation by identifying certain causal features of human perception as essential to making it reliable and identifying only those instruments that share these features as sharing the same (potential) degree of reliability as unaided perception ignores the fact that there may be other, dissimilar, ways in which the same degree of reliability can be attained. Finally, in focusing their definition of relevant similarity to human perception on the sorts of physical processes that occur between light interacting with an object and light being received by the retina, existing accounts of observation, suffer from the endpoint problem. Relevant similarity to human perception, if this approach is at least on the right track, must take the visual system to involve at least some neural mechanisms. Figure 2.3 illustrates this point. The endpoint problem consists in thinking that the potential observational status of imaging technologies (PM) is to be drawn by analogy to PH1. As the Chapter 3 will discuss, however, the analogy must be drawn with PH2.

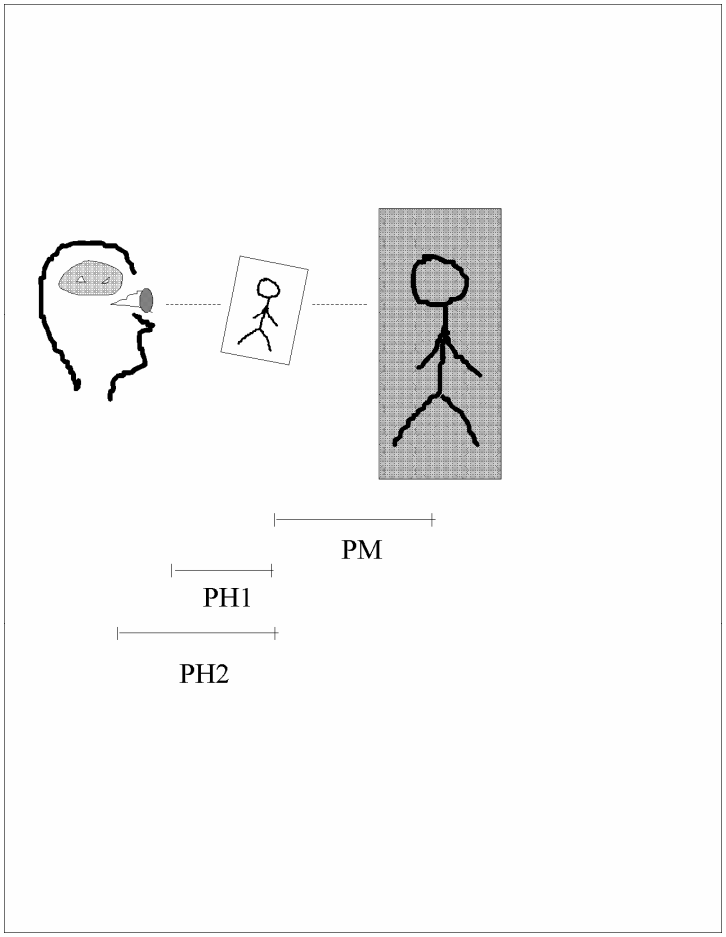


Figure 2.3 The scope of perception.

PM designates the processes involved in image production using a mathematized imaging technology. PH1 designates the processes involved in unaided human perception from the interaction of the object with light up to light hitting the retina. PH2 designates the processes involved in human perception including those neural mechanisms that form part of the visual system

While none of the accounts of observation discussed in this chapter is ultimately successful, they are each attempting to capture two features of observation that seem to be correct: that it must preserve spatial relationships and that this preservation must be ensured by a particular process. It is the nature of this process - and particularly its presumed reliability - that underlies the presumed epistemic privilege of observational evidence. The problems with existing accounts of observation follow from the fact that each is a benchmark strategy. I claim that the solution is to be found by supplementing a benchmark with a grounding strategy. In order to

justify any account of what relevant similarity to human perception is, we need to be clear on how relevance is to be determined. Are we concerned solely with physical similarity of processes? That might be plausible if it were not for the endpoint problem. Once we recognize that the human visual system, even if it is taken as pre-conceptual and strictly separable from cognitive processes,[51] involves more than retinal stimulation by light, an account of observation that is based on the physical properties of light (or a wider range of electromagnetic radiation) and causal processes involving light and light detection will be inadequate. The involved neural mechanisms must be accommodated and these do not involve physical interactions with light. But if the straightforward physics of signal and signal detection are inadequate, what determines relevant similarity? The idea that observation involves structure preservation produced by physical processes is based on the idea that (certain) physical processes are reliably able to transmit and conserve structural properties from the object the representation of it. The idea that all the processes in human perception are particularly good at doing this leads to the perception-reliability problem. Despite this, the basic intuition seems correct. A grounding strategy can take this intuition and attempt to define relevant similarity in terms of the sorts of features that make perception reliable – when it is – and their presence or absence in other imaging systems. Importantly, this may not exclude certain types of inference or mathematical or statistical processing. As the case of attenuation correction showed, there are cases where you can better preserve structural relationships by using inferential steps than you can by allowing only physical processes.

It should be noted that my goal is not to develop an account of observation that will accommodate PET. It may very well be that any account of observation that retains, as the anthropocentric interpretation of empiricism must, some non-epistemic criterion of similarity to

[51] As Pylyshyn (2003) argues.

human perception will exclude PET and/or other mathematized imaging technologies. However, it is not a necessary condition of a good account of observation that observation must hold a place of special epistemic privilege. The available options for the nature and epistemic status of mathematized imaging technologies were outlined in Table 2.1. The key possibilities are: 1) that epistemic privilege is always and specifically associated with observation, and 2) that non-observational processes can sometimes share the epistemic status of observational processes. For the anthropocentric empiricist to retain the central principle that the best source of evidence is human sensory perception (and allowable extensions of it), she must present an argument that identifies some epistemically relevant similarity between human perception and instruments that are included under the name of observation. It is not sufficient for her simply to claim that all epistemically privileged forms of data production count as observation. If, on the other hand, no such similarity can be found, then we must reject this interpretation of empiricism since there will be no connection between some instrument being like human perception and it being as good a source of evidence as human perception. The next chapter will examine whether any such relevant similarity can be found by using a grounding strategy that defines relevant similarity in terms of the reliability-making features of human perception.

3. Can imaging technologies be like human perception (and does it matter)?

3.1. Introduction

The last chapter showed that existing accounts of observation use benchmark strategies which assume that it is possible to define the scope of observation, i.e. to identify all systems that are potentially capable of producing data that preserves the relevant features of the observed object (spatial structure, color, etc.) to a sufficiently high degree, by reference to a benchmark that is held to be epistemically privileged. The anthropocentric interpretation of empiricism takes human perception (HP) to be that benchmark. If an imaging technology is relevantly similar to HP, the argument goes, it can produce similarly structure-preserving, reliable data and so share the epistemic privilege of HP. A real problem for this approach is how to understand what counts as relevant similarity. As the PET example showed, sometimes making a system seeminglyly less similar to HP by performing attenuation correction actually increases the degree to which the data preserves spatial features of the object under investigation. Thus, similarity to the human visual system in at least some respects is not required for reliability. This chapter will develop an account of observation that explicitly recognizes the potential for reliability and similarity to HP to occur either together or apart, an account that is encapsulated in what I will call the *grounded benchmark criterion* (GBC). The GBC, roughly stated, asserts that we can perceive via an apparatus if and only if the apparatus is similar to human perception (HP) with respect to those features that make HP reliable. If the stronger version of empiricism – according to which sense experience is not simply required for us to causally interact with the world, but has a uniquely high degree of epistemic privilege - is to be successfully defended, then the GBC must show that the reliability-making features of HP are uniquely associated with systems that

are similar to HP with respect to these features. If reliability can be obtained in other, non-similar, ways, then this version of empiricism fails.

The first aim of this chapter will be to clarify and explicate this criterion. The second aim will be to show that the reliability-making features that are identified using the GBC are not uniquely associated with systems that are similar to the human visual system. What matters for reliability is that the data preserve the relevant features of the object under investigation. This can be achieved in a variety of ways, many of which do not appear to be similar to HP. I say "appear" because part of the work of investigating the GBC will involve showing that our understanding of the human visual system and, in particular, of the things that contribute to making it reliable (when it is), is not yet sufficiently complete for us to make use of a benchmark strategy in order to assess the epistemic status of various imaging technologies. We simply don't yet know enough to be able to specify what features a system must have or not have to be relevantly similar to HP in its reliability-making aspects. If a benchmark strategy is to be a useful tool in trying to explain actual scientific practice, then it must not just specify that in order for instrument X to count as a mode of observation, it must be like HP in some relevant ways, but we must also be able to specify what those ways are in enough detail to be able to judge whether or not PET or some other instrument does or does not possess these features. Since we cannot do this, and since there is no independent argument to be made that generally HP-like mechanisms possess a uniquely high degree of reliability, we must reject the anthropocentric version of empiricism.

3.2. Some preliminaries

Before anything else can proceed, there is a matter of terminology that stands in need of clarification. In the previous chapter, I stipulated that I would use "observation" to include

unaided HP plus whatever a given author regarded as legitimate extensions of it. The above paragraph began by referring to accounts of observation and then switched to discussing perception. This was not an accidental oversight, but was quite deliberate. In discussions of empirical access to the world, "observation" is the term that is most commonly used[52] (Hacking 1983; Shapere 1982; Azzouni 2004; Kosso 1992; Machamer 1970; Menuge 1995). It is usually used to refer to a privileged mode of gaining access to phenomena and is associated both with a (perhaps false) dichotomy between things that are observable and things that are not and with an account of scientific knowledge that is based on observation sentences or observation reports. To introduce some consistency in the various usages of the authors whose work was discussed in the previous chapter, I have also used the term "observation" until this point to capture everything that is epistemically good in virtue of being like HP in some way. Since benchmark strategies characteristically define observation both as epistemically privileged modes of getting information and justifying claims about the world and in terms of likeness to human perception, there is little reason for them to distinguish between the terms "observation" and "perception" when what is at issue are the epistemic qualities associated with the class of things relevantly like human perception. However, I have claimed that this sort of strategy fails to define likeness in an epistemically relevant way—unaided human perception is taken as the benchmark for epistemically privileged ways of gaining access to the external world but extension of the term "observation" to various forms of technically assisted perception is allowed or disallowed according to physical similarities that are either not essential or not sufficient to guarantee that this extension is granted to all (and only) those systems that bear an epistemic similarity to HP.

[52] Hacking frequently refers instead to "seeing", though, interestingly, he does not use the term "perception". Outside of the philosophy of science – in particular, within epistemology and philosophy of perception – it is "perception" rather than "observation" that is normally used even when referring to extensions of the human visual system.

Whether or not some instrument involves causal processes that are similar to HP does not always determine whether it is capable of producing data that preserves features of objects in the world as well as does HP. Therefore, I want now to be able to distinguish between: a) methods that are reliable (in the sense that the structure of the relevant parts of the world are preserved in the data via the causal[53] processes involved in the method)[54] and get this reliability by using processes that are the same or similar way to those involved in HP, and b) methods that are reliable and get this reliability via processes that are different from those involved in HP. While benchmark strategies typically use "observation" to refer to the methods included in (a) and deny or at least ignore that (b) is not empty, I will instead use "perception" to refer to (a) and "observation" to refer to the disjunction of (a) and (b). Because I am interested in the epistemology of imaging technologies and there is no epistemic difference between (a) and (b), I will not coin a term to refer specifically to (b).

A second issue that merits a few remarks at this point is that of methodology. I have claimed that the causal and epistemic aspects of perception can be separated, at least conceptually, and that one might, for instance, want to define some system as causally like HP but deny it the epistemic privilege traditionally associated with HP (see Chapter 2, Table 2.1, box 3) or, alternatively, as causally non-perceptual but epistemically good (see Table 2.1, box 2). Why, then, propose a criterion that reconnects the causal and the epistemic aspects of perception? Why not move on directly to an account that focuses solely on the epistemic aspect and denies the epistemic relevance of causal similarity to HP? First of all, while the two aspects are separable in principle, it remains true that it is the causal processes that determine the reliability

[53] By "causal" I mean both straightforward physical processes like reflection of light and waves hitting a detector and mathematical and statistical processing. Though mathematical and statistical transformations are not causal in and of themselves, the application of algorithms can be interpreted as causing features of the data set to change.
[54] A more precise account of reliability will be developed in Chapter 4.

of a method since they are what either preserve or fail to preserve features of the world in the representation of it that is the data. Second, if there is to be any chance for the anthropocentric empiricist to retain the principle of epistemic privilege of our senses, then there must be some causal similarity to HP that links *all* forms of epistemically privileged data acquisition. While I have suggested that it is the epistemic aspect (understood here in terms of reliability) itself that is what matters in the end, it is not the case that an initial examination of specifically perceptual reliability is an unnecessary detour.

The second reason to specifically focus on perceptual reliability is that we need to do so in order to know if there is any principled reason to distinguish between the alternatives of identifying epistemically good processes and dividing those into two classes – perceptual and non-perceptual – or calling everything that reaches a certain threshold of reliability perceptual. While making this distinction doesn't serve an epistemic purpose on my account, it might still seem appealing in that there are some modes of reliable data acquisition that many people would greatly resist calling perceptual (e.g. instruments that count gamma rays that make it through the earth's atmosphere). To justify the claim that it makes no difference whether you call perception all of what I have termed observation - as long as it is epistemology that you are interested in – it helps to be able to show that, in terms of structure preservation, perceptual and non-perceptual reliability are just the same sort of thing, that they are not different but equal in the sense that perceptual reliability involves a distinct subclass of reliable processes.[55] In this way, it also serves as a connecting step from the anthropocentric empiricist's claim that perceptual evidence

[55] It should be noted up front, however, that I will be claiming that we have insufficient knowledge to establish the reliability of the sub-processes that contribute to HP. The argument for the reliability of HP must, instead, be made for the process as a whole and not based on an understanding of the precise processes that contribute to its reliability.

has a special privilege to the claim that it is reliable evidence, regardless of the sorts of distal processes involved, that is good.[56]

This, then, brings us back to the question of what the consequences are for empiricism if extensions of our perceptual capacities that are *not* connected to our native capacities in any causally significant way, are found to share the special epistemic status of perception. The idea that unaided and appropriately aided (bearing a relevant similarity to unaided HP) sense perception bear a uniquely high degree of epistemic privilege with respect to establishing claims about the natural world was central to the second, anthropocentric interpretation of empiricism outlined in Chapter 1. By establishing that the GBC is the best possible account that can be given of how the scope of observation should be extended—since it identifies epistemically relevant features of HP as these to which an instrument must bear similarity—and then showing that it fails to isolate *all* instruments that share this privilege, this version of empiricism is shown not to be viable. The first interpretation, according to which sense experience is required simply in order for us to have contact with the external world, can still be maintained. However, all this version says is that the proximal end of some series or chain of events that comprise an investigation of the world must be something that is accessible to our senses: data that is output by some instrument must be visible (or, less frequently, audible, tactile, etc.). This interpretation, as it stands, doesn't allow us to make any distinction between the epistemic status of various distal processes and so is not useful for the project of assessing mathematized imaging technologies unless it can be substantially elaborated. This elaboration will come in the next chapter in the form of an account of reliability, but the grounds for doing so will flow from the arguments of this chapter since the failure of the GBC shows that what the stronger interpretation

[56] Of course, this could also be achieved by starting with a general account of reliability and moving back to show that perception is just a specific case of this sort of reliability, defined in terms of causal likeness.

of empiricism was trying to get at was simply reliability. Reliability, then, is the basis for an expanded version of the first interpretation of empiricism, which, when fleshed out, becomes essentially a grounded empiricism as opposed to a benchmark or anthropocentric empiricism.

3.3. The GBC

The previous chapter identified three problems with existing accounts of observation. To review briefly, the problems are as follows:

1) The benchmark problem. The use of what I have termed *benchmark strategies* to define the boundaries of observation are inevitably forced to make an arbitrary choice of which aspects of HP are relevant for comparisons of likeness or similarity unless they also incorporate a *grounding strategy* that aims to define the relevant features as those that contribute to making HP epistemically good.

2) The perception-reliability problem. Evidence that HP, while usually reliable, is not always reliable is not taken into account, and HP and reliability are presumed to coincide.

3) The endpoint problem. The assumption is made that HP involves only processes that occur between the external object and the retina of the observer. Cognitive processes that are part of the visual system processes are uniformly ignored. These processes, however, are of a very different nature than those occurring between the retina and the observed object and their inclusion significantly alters how HP is understood.

Solving these problems begins with the recognition that HP has both a causal and an epistemic aspect. The problem of whether imaging technologies are forms of perception, then, also involves both a causal and an epistemic question. This is not to say that the causal story of perception is divorced from the epistemic story – in fact, just the opposite is the case– but rather that similarity to any or all causal processes involved in HP is not a necessary condition for

epistemic goodness. The assumption that such a connection exists is at the root of the perception-reliability problem, so showing it to be false will go a long way toward repairing this error. The distinction between the causal and the epistemic aspects also helps solve the benchmark problem since it facilitates the development of a grounding strategy by making it easier to identify the epistemically relevant and irrelevant parts of the causal story of HP. Since it is clear that (causal) cognitive processes play a major role in ensuring the epistemic qualifications of HP, the separation of the two aspects also helps to fix the third problem.

An account of perception[57] that aims to identify a class of epistemologically similar things cannot be a simple benchmark strategy, but must also incorporate a grounding strategy. A benchmark strategy alone places undue emphasis on the causal aspect of HP[58] and pays insufficient attention to the epistemic aspect. For any benchmark-type strategy to succeed, it must take the reliability-making features of human perception as the essential criteria for assessing relevant similarity to HP. This is to incorporate a grounding strategy—an attempt to justify the use of a certain benchmark (e.g. for good evidence production) by identifying that in virtue of which it is a suitable benchmark. This is the basic idea underlying the *grounded benchmark criterion* (GBC). Simply stated, the GBC says that *we can perceive via an apparatus if and only if the apparatus is similar to HP with respect to those features that make human perception reliable*. The GBC accepts that HP is, under appropriate conditions, a reliable means of getting certain kinds of information about the external world.[59] However, it recognizes that

[57] At least when perception is taken to be characterized in part by epistemic features, as it is must be by empiricism.
[58] While I am not aware of any accounts that do this, a benchmark strategy could also be based exclusively on the epistemic aspects of HP. This would have the effect of producing an account of what I am now calling observation. In other words, this might be how one would go about producing a general account of reliability (though this would not require that the starting point be HP).
[59] I will not be considering any kind of global skepticism here. If this problem is to be taken seriously, it will affect both unaided HP and human use of imaging technologies equally since there must be an end human user of the technology - or, more specifically, of data acquired using the technology - in order for there to be any question of knowledge.

HP is not infallible and that only the sorts of processes that contribute to its reliability are the ones that we want to identify as contributing to the epistemic status of HP.

A few points need to be made here. The use of the terms "perceive", "perceptual" and/or "perception" in this context will undoubtedly raise a red flag in the minds of some readers despite my earlier specification of what I will mean by *perception*. I want to be very clear, therefore, both about what I mean by these terms and about what I expect – and, just as importantly, what I do not expect - my account of perception to do. First of all, I do not in any way intend to be providing a universal, univocal definition of perception. This would not only be far too large a task, but would be a project doomed to failure. There are many different reasons one might have for wanting a definition of perception and many different functions to be served by such a definition. For instance, some authors (e.g. Keeley 1999, 2002) are interested in perception from a functional, evolutionary perspective and want to individuate modes of perception and distinguish them from processes that are merely detection. An important aspect of Keeley's (1999, 2002) account of sensory modalities is *dedication*: sensation requires dedicated anatomical structures to carry out specific kinds of sensory discriminations. Thus, merely because an organism can respond to electrical stimulation does not mean that it can *sense* electricity (it merely detects it). Perception in this context is not connected to the epistemic issues I am concerned with and can certainly be defined in a different way without introducing any sort of inconsistency. Secondly, I am not concerned with the question of whether machines can perceive. This question partially overlaps with mine insofar as the sorts of machine-based causal processes that can count as or enable perception are common to both, but I am not interested in whether machines can be claimed to have knowledge. I take it for granted that all imaging technologies (certainly all of those currently in use) have at least an eventual human

observer, user, or interpreter of the data and that any or all requirements that knowledge be a part of perception can be met in this way. What I *do* intend by the term perception is nothing more than is defined by the GBC: perception is a mode of acquiring data about the external world that both involves physical processes similar to HP and is able to make reliable discriminations (of color, shape, relative size, texture, etc.[60]) to at least approximately the same degree that the human visual system can. Perception, in other words, is anything that falls into Class I in Table 3.1. I will contrast perception with observation (data-acquisition processes that are reliable but may or may not bear any causal similarity to HP; Class III in Table 2.1), and with epistemically inferior forms of data acquisition, both like and unlike HP (Classes II and IV, respectively).

The notions of similarity and reliability are clearly crucial to the GBC but have been left very vague in all that I have said above. Much more will be said about them, both here and, in the case of reliability, in Chapter 4. Before moving on to that, however, I first want to present some simple cases to illustrate that the GBC provides us with intuitively correct answers for simple examples of all four types shown in Table 3.1. For this purpose, a common sense understanding of both similarity and reliability will suffice.

Reliability

Similar to human perception		**Yes**	**No**
	Yes	**Class I**	**Class II**
	No	**Class III**	**Class IV**

Table 3.1 Reliability and likeness to human perception.

[60] As I will discuss later in this chapter, it is not required that a perceptual process be able to make exactly the same sorts of discriminations as the human visual system. A system such as a phase contrast microscope, for instance, that cannot be used to perceive color, is not judged to be non-perceptual on that basis alone.

Falling into Class I are systems are those that satisfy the GBC and count as perceptual. They are both reliable[61] sources of data and similar to HP. HP, under appropriate operating conditions, is, of course, an example of this class. Another example is the light microscope (used by a human observer). The material composition and/or operating conditions of the light microscope are very different from those of the human visual system, but the light optics of the two are very similar and both generally display a high degree of reliability in allowing objects to be discriminated on the basis of shape, texture, relative size and position. Class II systems are those that are similar to HP but do not share its reliability. According to the empiricist and the GBC, this class has no members. Since relevant similarity to HP is supposed to be sufficient to ensure reliability—though it may not turn out to be necessary—nothing that bears the relevant similarity to HP can be unreliable in the appropriate operating conditions. Class III systems are those that are dissimilar to HP yet share with it a high degree of reliability. This is an important class since it may be hoped or expected to hold imaging technologies that fail to meet the GBC, and so do not fall into Class I. Such examples, however, are very complex and cannot be assessed using our intuitive ideas of similarity and reliability. A simple example of a system that would fall into this class, then, might be a stethoscope used to observe heart sounds and identify certain kinds of cardiac defect such as heart murmurs. Finally, Class IV systems are those that are both unreliable and dissimilar to HP. There would seem to be little use in purposely designing an unreliable instrument, so members of this class will consist primarily of those that either have broken down, are used for a purpose or in a situation other than that for which they were intended (for which they may or may not be reliable), or which, though unreliable, are the

[61] Later, I will specify that they must possess both reliability and validity in the sense of those terms that is common in statistics and psychology, wherein a valid instrument is one that measures the intended phenomenon and a reliable instrument is one that yields repeatable results. When referring to HP itself, however, the concept of validity seems misplaced so I have chosen to leave it out of the simple statement of the GBC. Within epistemology, at least, the term reliability is often used in a sense that subsumes validity.

best or only available way for gathering some sort of data about a particular phenomenon. The use of a stethoscope to detect the presence of a healthy, functioning heart in an organism with an open circulatory system (such as a mollusk) would be an example of the second kind.

Though all of the above examples are quite simple, it should already be clear that the intuitive notions of similarity and reliability are being pushed to their useful limits (or perhaps beyond, in some cases) in explicating how these examples measure up to the GBC. It is time, then, to develop a more substantial account of similarity. A fuller exposition of reliability will be presented in the next chapter. For now, it can be fleshed out slightly by specifying that it is described by the two features that the previous chapter identified as needed in any account of perception: 1) it must preserve information about structural features of the perceived object or phenomenon (these will, of course, vary according to the perceptual modality in question, but will include such things as size, shape, color, etc.), and 2) this information must be preserved via the causal processes that make up perception.[62]

3.4. Similarity

In our everyday experience of the visual world, we very readily identify objects that are similar in color or shape. I can pick all of the blue marbles or all of the big ones out of a bag containing a collection of marbles in various colors and sizes. We also have great facility with identifying whether a pair or set of concepts is similar (Goodman 1955; Gentner, 2000). But just what is it that we are identifying when we identify two things as similar? The blue marbles are similar to one another in terms of one property - their color. They may or may not be absolutely identical in terms of their color- they may have differences in tone or hue. Some of them may

[62] More will need to be said about what sorts of processes are causally efficacious. In particular, whether and how the mathematical and statistical processes involved in many imaging technologies can be thought of as causal processes is problematic. While mathematical or statistical operations might not be causal, there is a sense in which the application of algorithms to data is causal. For now, I will use the term "causal" to include such processes.

also be similar to each other in terms of size, whether big or small. They might also be made of different materials with different densities or textures. If I were given a large blue glass marble and asked to pick out all the similar marbles from a bag containing a mixture of colors, sizes, and materials, then, how should I approach the task? Most likely, I would search for other large blue glass marbles and pick out only them. I would pick out, in other words, those which had the greatest number of properties in common with the target marble. It might be that the other marbles were absolutely identical to the target marble. To be all but numerically identical to another object is certainly to be similar to it. But just as certainly, this degree of similarity is not required. Similarity is not absolute but comes in degrees. It may also be relative to context. If all the marbles in the bag are various shades of blue, I may be inclined to pick out only those that are large, glass, and a shade of blue that closely matches that of the target. If, on the other hand, most of the marbles are orange, red, yellow, and brown, and the blue marbles are of a shade quite different from the target (say, the target is navy blue and the marbles in the bag are more the color of a robin's egg), I might pick out those of the robin's egg blue marbles that are large and glass even though I would have left them behind in a bag filled with many darker blue marbles.

In addition, what counts as relevant similarity is always relative to some set of interests. Suppose that there are no other marbles in the bag that are large, blue, and glass. Some are large and blue but not glass, others are blue and glass, but small, and others are large and glass, but other colors. Still others have only one property in common with the target marble: they are large or glass or blue. Should I now pick out all the blue glass marbles? All of the large blue ones? All of the blue marbles regardless of size or composition? Without further instructions, it is not clear how I should proceed. The marbles that have two properties in common with the target might be quantitatively more similar to it, but (assuming that all property matches are

weighted equally) is any of the three property pairs <blue, glass>, <large, blue>, or <large, glass> more similar than the others? And is it always the case that having more properties in common makes for a higher degree of similarity? If the purpose of picking out marbles that are like the big blue glass one is to get a collection of objects that can be crushed into glass powder, then the relevant similarity involves only one property and I should pick out all and only glass marbles, paying no attention to whether or not they have any other features in common with the target marble. So, on this informal account, similarity seems to be related to the properties of objects (if both the base and the target for assessment are objects) and is relative to both the context in which similarity is judged and to some purpose or set of interests. What more can be gained by turning to a more technical account?

The primary way in which similarity is understood more technically is in terms of isomorphism. Formally, an isomorphism is a bijective map f such that both f and its inverse f^{-1} are homomorphisms (structure-preserving mappings). Unless we are prepared to provide mathematical descriptions of all the objects and phenomena involved in HP and imaging technologies, this formal definition won't be of much use. However, some philosophers (e.g. Weitzenfeld 1984) have made use of a less formal definition of isomorphism. Less formally, an isomorphism is a map or relation between structures where each structure consists of a set of elements and a set of relations among those elements: "The word isomorphism applies when two complex structures can be mapped onto each other in such a way that to each part of one structure there is a corresponding part in the other structure, where "corresponding" means that the two parts play similar roles in their respective structures." (Hofstadter 1979, 49). A very similar description is given by Hacking, though without any mention of isomorphism: "What is a good map? After discarding aberrations or artifacts, the map should represent some structure

in essentially the same two- or three-dimensional set of relationships as are actually present in the specimen." (Hacking 1983, 208). What seems to be crucial here is the idea that it is structural similarity that matters and that this is assessed at some level of abstraction, not with respect to the identities of the elements themselves. If there exists an isomorphism between two structures, then they are "the same" at some level of description. For example, the small blue glass marble and the large red metal marble are both spheres. Their material composition and color are different, but their geometric structures are isomorphic. This less formal use of isomorphism is helpful, but isn't sufficient on its own to fully characterize the notion of similarity that in needed in the GBC. The crucial problems that remain unanswered are: 1) which elements or features of these systems need to be included as elements or relations of HP or imaging technologies in order to identify epistemically relevant similarities, and 2) what level of description is interesting and informative for the same purpose.

To begin with the first question, there are three general ways in which we might attempt to understand similarity of HP to an imaging system: 1) in terms of the material composition of its parts, 2) in terms of the inputs and outputs of the systems, and 3) in terms of the types of mechanisms involved. The first of these seems to be obviously wrong. Why should the physical stuff out of which an imaging system is made matter? If we were to restrict the material composition to the sorts of proteins, etc, that make up the human (or even vertebrate) eye, we exclude many other classes of animals, especially those with compound eyes from having perception. This is not necessarily a problem since the GBC is concerned with identifying not just causal processes that are involved in seeing, but those that contribute to its epistemic goodness. Whether or not animals can have knowledge,[63] no one would deny that they at least

[63] Some such as Dretske (1969) have taken the impossibility of animal knowledge given a particular account of knowledge to be a serious argument against it.

have something (we might call it sensation) that is causally like HP but epistemically unreliable (i.e. that would fall into class II).[64] This is still possible by insisting that material similarity to HP be included in the GBC. However, to my knowledge, no author other than van Fraassen wants to deny that the light microscope can be used to observe (or in the terminology of this chapter, to perceive) – that this is an instrument that ought to fall within class I. And certainly the inorganic material that makes up a light microscope is very dissimilar to the organic material of the human visual system. If we were to make a serious attempt to include material composition as epistemically relevant, therefore, we would either be forced to go to a ridiculously abstract level of description - that the system has to be made of material stuff – or rework the description of the material composition in functional terms – e.g. that (some of) the material has to permit the passage of electromagnetic radiation in the visible wavelength, that lenses with certain functional properties be present, etc. The former is uninteresting and useless when it comes to grounding the epistemic aspect of HP while the latter will be dealt with separately as explicitly in terms of mechanisms.

The second possibility – the input and output of the system – is more promising. Hacking's (1983) account, in fact, makes a quite successful case that the input and output of an imaging system and their similarity to HP matter for whether or not we can claim to be able to see using a particular instrument. He argues first of all that the input to a system need not be restricted to light within the visible spectrum, but that an instrument whose input is any sort of electromagnetic radiation, or even waves more generally, can enable us to see. Because the optical characteristics of light (reflection, refraction, etc.) are essential to the eventual production of the image, and, in particular, to the production of images that are "good maps" in that they have the desired sort of structural similarity with the object they represent. There remains some

[64] If they do have knowledge, of course, it must be epistemically reliable for at least certain applications.

question about how we should understand the output of HP as well as the inputs to various sorts of imaging technologies (is the input some property of the phenomenon we set out to examine or instead what is detected by the detector system that is a part of the imaging system?), and these will be discussed in section 3.4.1. For now, however, it seems that the input and output to a system are causal elements that do play an epistemic role.

The third way in which we can characterize HP – in terms of function – will also be very important. By function I mean something like how the components of a system interact to perform the various types of processes that take a given input and produce a certain output. Hacking's account of why any sort of electromagnetic radiation can permit seeing already includes some reference to the functioning of the apparatus – how features of electromagnetic radiation are involved in or require specific processes to be performed by the system (e.g. diffraction of light requires that diffracted rays be at least partially recaptured in order to produce a good map of the specimen). It is within the domain of function that the cognitive portions of HP as well as the mathematical and statistical processes included in assorted imaging technologies will need to be accounted for. This, then, will be the most important area in which to connect causal to epistemic features of perception. It will also be the most difficult one in which to try to identify an interesting and informative level of description for processes whose physical instantiations and lower level functional descriptions are very disparate. A further complication here is the fact that many aspects of the functioning of HP are currently unknown. A great deal is known about which areas are involved in certain types of visual processing, but much less is known about how this processing is actually accomplished. Given these significant gaps in our knowledge, the best that can be provided at this time is a relatively abstract

description of the sorts of processes that HP must be able to perform in order to ensure the degree of reliability that it has in for certain tasks.

To try to provide anything even vaguely approaching a complete functional description of the mechanisms making up the human visual system would be a daunting and ultimately impossible task. It would be impossible in anything less than a multivolume book not only because of the vast amount that is known about the visual system, but even then because of the amount that is currently unknown. Fortunately, for the purpose of my argument, the extent of our current ignorance does not present a problem. If we do not know enough about the workings of the human visual system to provide a complete account of how it achieves its function(s), then it is instead the defender of empiricism who faces a challenge. For once we explicitly acknowledge the extent and variety of neural processing that is required to get veridical perception, we must also admit the possibility that there are instruments that perform relevantly similar processing operations and that should, based on any sort of benchmark criterion (including the GBC), be taken to share the epistemic status of HP. If the epistemic goodness of an imaging technology is, for the empiricist, ultimately connected to its similarity to HP, then her argument for or against some mathematized imaging technology must appeal, in part, to the similarity or dissimilarity between the computations performed by the human visual system[65] and the imaging technology in question. At present, this part of her argument simply cannot be made convincingly. This is not to say that the argument will not be able to be made at some point in the future when the state of knowledge about the human visual system is more complete, but surely we need not wait until that date in order to assess the epistemic status of imaging technologies. The empiricist argument is not the only one by which to establish the reliability of

[65] I do not mean by the use of this phrase that cells are literally performing computations, but rather in the sense identified by Grush (2001, 156) that groups of neurons compute in the sense that they process information as if they were implementing computable functions.

an imaging technology. That, however, is the still distant conclusion of this chapter. Before we can get there, it remains to be demonstrated that an argument based on any benchmark criterion, even the GBC, must fail.

3.4.1. Similarity of Input and Output

The input to HP is straightforward: it is light within the visible range. It is true that there are different types of photoreceptors - rods and cones - as well as differences among the cones, but these are functional differences that have to do with the responsiveness of the different cell types to a specific input. If we take input to an imaging system to simply be the entities that that causally interact with the detection system – here, the photoreceptor cells of the retina – then all that matters is that it is light in the range of wavelengths from 400-750 nm that can be detected. Questions about light intensity, pattern of light, and particular wavelengths within the visible spectrum have to do with functional properties of HP. The inputs to PET and CLSM are similarly straightforward: in the case of PET, it is 511 keV photons, in the case of CLSM, it is laser light of specific wavelengths.[66] It is important to be able to separate the irradiating laser light from the fluorescent light emitted by the sample and to separate the fluorescent light into various regions of the light spectrum when multiple labels have been used. This is achieved through the use of dichroic mirrors and optical filters to separate light of different wavelengths.[67] The light is then detected, in the case of CLSM, by photomultiplier tubes (PMTs). The input to the microscope, then, is light within the visible spectrum.[68] It might be objected that the input in this case should be taken to be only light of particular wavelengths - which ones being dependent

[66] Depending on the fluorochrome that was used to label the sample, a different wavelength of light will be required to excite the fluorochrome, causing it to fluoresce – to emit light in a particular wavelength. These are referred to, respectively, as the excitation and emission wavelengths.

[67] Each imaging channel has its own PMT. Thus, for instance, if a sample is dual-labelled with a fluorochrome that emits light in the green range and one that emits light in the red range, a dichroic mirror is used to split the longer wavelength (red) light from the shorter wavelength (green) light, each of which is directed towards different PMTs. This is crucial in order to collect data in different channels separately since the output of each PMT is determined only by the total light hitting it – the wavelength doesn't matter.

[68] Some fluorochromes emit within the UV range and can also be detected, but most emit within the visible range.

on the fluorochromes used. This is not the case, however, since what matters with respect to detection is simply the amount of light that hits the PMTs. It is the other physical components of the microscope – in particular, the wavelength of light emitted by the laser and the use of dichroic mirrors to split emitted light – that determine which wavelengths will be received by the PMTs. While any particular use of the microscope will not involve the PMTs detecting wavelengths across the spectrum, neither does HP always involve detection of the full range. If, for instance, I were in a closed room with plain white-painted walls, illuminated with pure green light (say, of 518 nm), and my head restrained so that I could look only at the wall (not at my own body or clothing), I would see only light of a particular wavelength. It would not be true to say, however, that the input to my visual system overall is limited to this green wavelength. Due to the particular physical situation in which I find myself, that may be the only actual input to my visual system at the moment, but in a different physical context, my same visual system can detect a much broader range of wavelengths. It is the in theory input, not the input in a specific application of the system, that is relevant for the purpose of defining the input to the system in general.

So how do these inputs bear on the reliability of HP and so fit with the GBC? As far as input is concerned, what matters is that there should be a reliable transmission of the input entity (light, high energy photons) both between its source and the light (photons, etc.) emitted/reflected/refracted etc. by the object that is emitting/reflecting/refracting/otherwise interacting with it and between the object and the detection system. This is what Hacking (1983) correctly sought to capture with the idea of physical processes and the argument that any electromagnetic radiation behaves in similar ways and should be allowed to count as permitting "seeing". There are additional complications introduced when the phenomenon that we are

interested is only distantly related to the input to the imaging system, and these will be addressed in Chapter 4. For the purpose of judging the immediate inputs to the imaging system to be similarly reliable, however, these complications need not be considered.

While identifying and comparing the input to HP and imaging technologies is relatively straightforward, the same is not true of outputs. In general terms, the output ought to be taken to be what the observer *uses*. In the case of HP, this could be taken to be phenomenal experience, qualia, or some feature of the visual system of the brain. What the output should be taken to be, then, will be at least partly relative to the purpose to which the observer is putting the observation. Philosophers have long debated the nature of visual representations. The situation is perhaps less contentious, but still unclear when one looks instead at the scientific literature. Specific patterns of neural activity in and across various levels of the hierarchical visual pathway somehow are correlated with objects (visually accessible properties of the external world) and there is good evidence that certain cell types and certain cortical areas represent different features[69], but just how it is that the physical states of groups of cells and the relations between them can be taken to represent objects is unclear. We just don't yet know what, in neural terms, corresponds to the final representation of what we see. The neural correlate may be the activation of particular cells, of particular cell assemblies, of particular temporal patterns of activation within or across cells regardless of which cells display this pattern, or it may be a combination of spatial and temporal patterns (Treisman 1999). Moreover, since mental representations, whatever they turn out to be, at a minimum do not have spatial characteristics whose structure can be compared with the object being perceived.

[69] Treisman (1998) identifies six different types of object representation, each specialized for a different task and encoding different properties with varying degrees of detail.

While computational models are often taken to be indispensable for understanding representation,[70] it is not clear that computational models will tell us what we want to know about how the actual system. Grush (2001) has argued that computational neuroscience is unable to provide an explanation of why some neural states but not others are representations and of what the content is of those neural states that are representations. However, we need not be overly troubled by this conclusion if we share in Grush's optimism that it is possible to formulate a theory that can provide such explanations. If this is the case, then this simply counts as one more area in which we do not have sufficient knowledge to adequately fill out the sort of account that is needed for the GBC or any benchmark strategy to succeed.

Similarly, though less obviously, what we should take to be the output of PET and confocal microscopy, is not self-evident. Though the output images are usually brightly colored images, the fact that the input is digitized early in the process and that the end result of the extensive statistical processing is an intensity value for each voxel which can then be displayed as a 2-D or 3-D image by assigning a color to specific ranges of intensities suggests that it is not unreasonable to take the final output to be the array of numbers corresponding to each voxel. For the question of reliability, it makes no difference which we take to be the 'true' output since the conversion from the numerical data to images is unproblematic. The question will arise again in Chapter 5, however, when we examine what epistemic value might be gained by producing images.

[70] "Anatomical and physiological data provide important clues about form vision, but a complete understanding of visual function requires us to work backwards from the neural implementation to the computational principles involved. Experiments describe the tuning profiles of cells in a particular area and, in some cases, the anatomical distribution of cells with similar response properties. Computational considerations must be invoked to address the underlying representation and its functional role in vision." Gallant (2000, 324).

3.4.2. Mechanisms[71]

Before attempting to break human visual perception into a limited number of functions for which mechanisms can be sought, it will help first to be clear about what the overall function of the visual system is. Some prominent researchers in the field of vision have described it as follows:

> "The overall goal of the visual system is to provide an accurate three-dimensional description of the environment over time which includes the identification of objects and events." (Geisler and Albrecht 2000, 121)

This description[72] captures several key points that should be kept in mind: 1) the visual system is supposed to provide accurate representations, 2) in order to achieve this, it needs to represent the world in its full three dimensions, despite the fact that the retinal display is only two-dimensional, 3) vision is temporally extended, and 4) correct individuation of objects (and events) is crucial. It should also be kept in mind that this overall function is achieved only by the visual system taken as a whole – from the retina through the primary visual cortex and higher visual areas.[73] We need, in other words, to include both well-characterized parts of vision such as the response patterns of rods and cones to light of various intensity and wavelength, and less well understood aspects such as the influence and control of attention. It will also be helpful to

[71] There has been considerable recent philosophical discussion about mechanisms, but differences between, for instance, Glennan's (1996, 2002) and Machamer, Darden and Craver's (2000; Machamer 2004) accounts will not affect this discussion. Here, the term is meant to track the scientific sense without being specific about its philosophical underpinnings.

[72] This description could be amended to explicitly recognize that not all of human vision is visual perception. The simplest kind of sensitivity to light is photosensitivity – the ability to detect different intensities of light within the visible spectrum via photosensitive molecules. This ability is present in many organisms, including even many single-celled organisms, and is important in humans for its role in regulating circadian rhythms (Shepherd 1988, 326).

[73] I do not mean by this that *each* aspect of vision involves *all* visual areas. The visual system is characterized by hierarchical processing and functional specialization and different functions are carried out in different higher visual areas. For instance, the middle temporal, or MT, area is both necessary and sufficient (among higher visual areas) for perception of motion (Farah 2000, 45).

recall that the goal here is not to provide a full functional account of the human visual system, but to examine what functions contribute to the reliability of HP. Moreover, since the final goal in describing this set of structure-preserving functions is to assess whether or not we can argue that imaging technologies such as PET are relevantly similar to HP with respect to these functions, we need not consider any functions of HP that have no analog in imaging technologies. This means that we must consider how the visual system produces representations of object tokens (a viewpoint-relative representation of an intact object as it is currently seen) but need not take into account how the visual system produces representations of object type (recognition of an object as a member of some category of object – e.g. as a cat) or representations that are based on further knowledge associated with that category – e.g. that it is warm, purrs when pet, and likes to chase birds). Finally, it is important to recall that the claim being made here is not that we can use the GBC to validate imaging technologies, but rather that we cannot do so (though I will go on to argue, briefly here and at more length in the next chapter, that possession of reliability-making features has no connection with bearing relevant similarity to HP). While the claim that we *can* justify PET and other mathematized imaging technologies by using the GBC and arguing that they are relevantly similar to HP with respect to its reliability-producing features requires that a complete account be given of what these features are, the negative argument requires only that it be shown that no such account can be provided for HP. Thus, in what follows, I will not attempt to provide a complete description of any of the functions of HP. While it may be of some interest to consider, for instance, whether or not some mechanisms of attention or top-down processing at various levels of the visual system should be considered to blur the line between observation and theory in a way that should make the

empiricist uneasy, such issues will not affect my conclusion here and will be set aside for examination at some other time.

With those constraints in mind, I propose that we should take the visual system to have as the following three general functions:

1. Selective reduction of information available to the system.
2. Transmission (which will include transformation) of information from input through various levels of processing to the final output.
3. Association of separately processed information to produce a unified representation of objects (one in which each visual feature of an object, such as its shape, is correctly associated with each of its other features, such as color and location).

Each of these functions must be achieved in order that the final output of HP (or an imaging technology) is reliable. The second and third very obviously serve this end, but the first may not seem quite as clear. Any elementary psychology text, however, will point out that the brain is limited in its processing capability and that the (presumably evolutionary) solution to this is to fully process only the most relevant stimuli. Why and in what sense its processing capacity is limited, however, is less clear. Farah (2000, 175) identifies two sorts of processing limitations. First, there is a response bottleneck in the sense that our behavioral responses (e.g. reaching for objects or moving our eyes towards stimuli) are limited - we cannot respond simultaneously to large numbers of stimuli - so it may be helpful to limit the number of stimuli within the visual system itself. Second, the use of distributed representations[74] within the visual system suggests that a common set of neurons will represent a variety of different stimuli and when more than one stimulus is represented simultaneously, problems with mutual interference and crosstalk are

[74] Readers unfamiliar with the visual system should refer to the appendix for a brief description of its organization.

introduced. Selective reduction of information that gets fully processed by the visual system, therefore, helps to ensure its reliability. This function is carried out largely by the retina. The retina makes substantial "front-end" reductions in the amount of information that gets transmitted to downstream parts of the visual system. It does not encode and transmit all of the information available in the retinal image; it is selective to a limited range of wavelengths, luminances around the mean, and temporal frequencies. Furthermore, over most of the visual field (with the exception of the fovea), the retina encodes a limited range of the available spatial frequencies. Within these broad limits, there are subpopulations of neurons that are even more selective, subdividing the information into narrower ranges. However, this function is not exclusively carried out by the retina; there are additional *attentional* mechanisms that select a small subset of the stimuli for extensive processing and consign the rest to only limited analysis. There remain many open questions regarding attention and at least six different cognitive models of attention have been proposed (see Shipp 2004), so the undisputed fact that attention does play an important role in limiting the amount of information that gets processed already indicates that this function is far simpler and far better understood in the case of PET and CLSM than in HP.[75] Farah (2000, 178-9) identifies two different sorts of questions, the first having to do with how attention affects processing at various levels in the visual system, the second concerning how the attention itself is controlled. With respect to the first question, while there is clear evidence that attention modulates processing in the extrastriate visual cortex, there is conflicting evidence for whether or not attention plays a role as early as the primary visual cortex.(for a review of multiple lines of evidence, see Farah 2000). In addition, the mechanisms of selection (within both early and late visual areas) are poorly understood. They appear to involve both facilitation

[75] This is not to say that the retinal mechanisms are completely understood, but just that our understanding of attention is much farther from being complete.

of attended stimuli and inhibition of unattended stimuli, but there is very little information on how such modulation of neuronal activity is achieved (see Hillyard, Vogel, and Luck 1998 for a discussion of alternative models; see Wolfe and Horowitz 2004 for a discussion of the attributes that (may) guide the deployment of attention). With respect to the second question, there is considerable, though not unambiguous, evidence that both the prefrontal and parietal cortices play a role in top-down attentional control (see Farah 2000). Again, however, the mechanism or mechanisms by which such control is exerted are unknown.

The situation is again much simpler in the case of PET and CLSM. Excess of information is, in fact, not a problem for PET. There the real difficulty is getting enough information. The signal-to-noise ratio is very poor, so data must be collected over a sufficient time period in order that statistical processing techniques that model and seek to eliminate noise from the data set can be applied. CLSM is also different in that the potential for excess information (which might be considered noise) is present. However, the confocal optics were specifically designed to eliminate all but the in-focus light from reaching the PMTs. In this case, then, information reduction is carried out by the optical components of the system and is very well understood.[76]

The second function, reliable transmission and transformation (e.g. from electromagnetic to chemical form) of information, is variously well and poorly understood for different aspects of HP. We have a good understanding of the behavior of light and this allows us to account for transmission of information between the object and the retina. We also have a good understanding of both molecular neurobiology and neurophysiology at the level of the single neuron. Tract-tracing has also provided us with fairly good information about the neural connections between various brain areas (see Rockland and Pandya 1979, Scannell et al. 1995).

[76] While this applies to CLSM, it is not true of confocal microscopes that use Nipkow disks. In this case, it is not fully understood how light projected through these spinning disks covered with an array of "pinholes" is able to achieve the same effect.

What we do not have is a good account of the functional architecture and this is crucial for knowing exactly how information in various processing streams gets transmitted and, in particular, modified by other inputs.

The biochemical mechanisms that occur at synapses and the molecular events that affect the membrane potential are, though not understood in every detail, relatively well characterized (for details, see Kandel, Schwartz and Jessel 2000). Interactions between neurons are fundamental to how information is exchanged within and between areas (or computing elements) of the brain. It is important to understand, for instance that when a synaptic terminal receives an all-or-nothing action potential from the neuron of which it is a terminal, it releases a chemical neurotransmitter that crosses the synaptic cleft and produces either depolarization or hyperpolarization in the postsynaptic neuron by opening particular ion channels. Summations of a number of depolarizations (excitatory inputs) within the time constant of the receiving neuron (typically 15-25 ms) produces sufficient depolarization that the neuron fires an action potential. The action potential is then conducted in an all-or-nothing manner from the axon through to the synaptic terminal where it can affect other neurons. Any inputs that the neuron receives that cause it to be hyperpolarized move the membrane potential away from the critical threshold at which an action potential is initiated and thus are described as inhibitory. The neuron can thus be thought of in a simple way as a computational element that sums its inputs within its time constant and, whenever this sum reaches a threshold, produces an action potential than propagates to all of its outputs. However, what is crucial for the transmission of information about aspects of the visual scene is not interactions between single neurons, but the distributed patterns of activation across large numbers of neurons. This requires knowledge not only of the structure of structural connections within and between brain areas, but of functional connections, a challenge that

remains unsolved. [77] It also includes knowing how to best model the summation of inputs, since in the absence of a good functional architecture, it isn't possible to assess exactly which and how many neurons are involved in a specific task and what their input and output cells are. As an additional complication, there is evidence that not all response patterns are linear (see, e.g. Edelman 1999). Thus, to try to understand how information about particular aspects of the visual scene gets transmitted though the groups of neurons in the various hierarchically organized areas, it is necessary to use computational models. There are very different approaches that can be taken to modeling the interactions between multiple neurons. One type of approach is to assume that the responses of the neurons are combined using a simple operation such as simple summation, Minkowski summation, or "winner-take-all". Another common approach, however, is to treat visual perception as Bayesian inference and use model particular visual tasks by using a so-called ideal observer who computes the most probable interpretation of the retinal stimulus by extracting all of the information[78] available in neuron responses and combining it with information about stimulus probabilities (for a review, see Kersten et al. 2004). The details of these approaches are not important here; what matters is that we simply do not have anything approaching a complete knowledge of the neural connections that are involved in specific visual tasks and that, in the absence of this knowledge, it is not possible to give an account of how information about specific visual features is transmitted through the various brain areas believed to be involved. Understanding how transmission of information occurs within and between individual neurons (the molecular neurobiology and neurophysiological accounts) is not an account of how information about color, shape, motion, and other aspects of the visual scene get

[77] There is a great deal of information on neuroanatomical structure, but the degree to which individual areas are interconnected (some areas connect with 8-10 others) has made it very difficult to establish a functional model of cortical architecture (see Vezoli et al. 2004).

[78] I am using "information" here in a strictly non-technical sense and not, for instance, the sense of Shannon-Weaver.

transmitted. To accomplish that, we need to know more about the patterns of interaction between groups of neurons, and that information is not currently available.

Contrast this with the case of PET where transmission of information can be described much more completely. As with HP, description of the features and behavior of photons is straightforward. After the photons are registered by the scintillation counter, however, the information literally is digital information stored on a computer. The statistical and mathematical computations that will be performed on it are literally computations, not simply capable of being described as or represented by computations. Furthermore, even if they are not necessarily perfect from the perspective of avoiding error,[79] the computations that are applied to the information are fully known. The various software packages that determine them[80] were written by computer scientists and the code contains full information about the processing that the initial count data can be subjected to. The same is true of CLSM. As an optical system, the first part of the story about transmission of information is, as with HP, to be found in our understanding of the properties of light. After light hits the various photomultiplier tubes, it is converted to digital form and stored on a computer as was the case with PET. Further transformations of that digital information again is performed through the use of various software packages and is exactly analogous to PET in the completeness of our knowledge of these transformations (should we choose to examine them), if not in the identity of the transformations.

The third function, association of separately processed information to produce a unified representation of objects, is related to the second in that it clearly involves transfer of information, but it also requires coordination of information. The evidence that separate, parallel

[79] This issue will be examined in the next chapter.
[80] Or, more accurately, that determines the set of processes that are available for use. There is, of course, room for the user to adjust settings and use or ignore certain options provided by the software.

processes are responsible for different aspects of vision and that specialized areas code represent different aspects of the visual scene, raise one of the biggest unsolved problems for our understanding of the visual system: how do we get from dispersed brain representations of different aspects of the world to the coherent, unified percepts that we experience? This is referred to as the *binding problem*. There are three aspects of the visual system that contribute to the possibility of binding. The first is the general one noted above that various properties of objects appear to be separately analyzed by specialized visual subsystems. Thus, while information from the same spatial location is implicitly bound by the cells that respond to it initially, at later stages the information from these cells appears to be routed to different neural populations, forming a distributed representation of an object's various properties. The second is that receptive fields at higher levels are large enough – up to 30^o in temporal areas – to generalize across a wide range of locations. Because visual scenes typically contain multiple objects, the question of which features belong to which objects could frequently arise. Third, coarse coding of different stimulus dimensions creates representations that may depend on ratios of activity in neurons with different but overlapping tuning. Whenever the perceptual representations of simultaneously present objects depend on distributed patterns of firing in populations of cells, the risk of superposition ambiguities within the same neural network will arise, creating a need to identify and to signal which units belong to the same representation.

Several different cognitive models for binding have been proposed. Prominent among them are Treisman's Feature Integration Theory (FIT) (Treisman 1982, 1998, 1999), Wolfe's Guided Search model (Wolfe, Cave, and Franzel 1989; Wolfe and Bennett 1996), and Reynolds and Desimone's biased competition model Reynolds and Desimone 1999). The nature of the differences is not crucial to the argument here and will not be described. The interested reader is

directed to Treisman (1999) for a summary of the disagreements between the models. The crucial point is that there currently exist multiple accounts of how binding is achieved and the available evidence is unable to distinguish between them. Furthermore, although there are differences between them, all suggest that attention plays an important role in binding. Thus, uncertainty about mechanisms of attention limits our ability to specify precisely how binding occurs, even if there were agreement on other aspects of the binding problem.

Though how this function is reliably achieved is still unclear in the case of HP, it is again less problematic in the case of PET and CLSM. In these instruments, the binding problem simply does not occur so there is no need for it to be solved in order for the output image to possess all of its own features. Of course, it is true that PET and CLSM do not represent all of the features of objects that HP does: neither can represent color (recall that the color in the images is artificial, corresponding to the intensity value calculated for each voxel rather than corresponding to the actual color of the object), nor can they represent motion as an independent feature of objects (though CLSM can represent motion as a consequence of its ability to detect change in position over time). However, the key aspect of PET and CLSM that allows them to avoid the binding problem is that they do not have separate processing of different object attributes. All that is measured is the light intensity or number of photons that hit the PMTs or scintillation counters when the instrument is scanning a particular location in the object (recall that both PET and CLSM involve scanning successive layers of the object). The ability to recover spatial information from this data is ensured by the physical arrangement of the detectors and (in the case of PET) the use of reconstruction algorithms. Thus, there exists only the need to ensure reliable transmission and transformation of the data, not an additional need to re-attach different features of objects.

3.5. Conclusion (what does the GBC tell us?)

The GBC states that we can perceive via an apparatus if and only if the apparatus is similar to HP with respect to those features that make human perception reliable, where the conclusion that some apparatus allows us to perceive is intended to establish that the evidence gathered using the apparatus shares the privileged epistemic status of HP. It was intended to solve difficulties with other benchmark strategies by both explicitly including the neural components of the visual system within the domain of required similarity and by providing a non-arbitrary way to decide the question of what counts as relevant similarity to HP for the purpose of extending the boundaries of epistemically privileged observation. It was, in other words, intended to provide the last best chance for the empiricist to justify the claim that instruments of various types are epistemically trustworthy if and only if they are an extension of our own senses. The basic idea underlying benchmark strategies (as employed by Hacking, Van Fraasssen, and Shapere) is that: 1) perceptual knowledge, though not indefeasible, is usually a very good way to get reliable information about the physical world, and 2) the way to be sure if the ways we have of extending our sensory capacities provide similarly reliable data is to see if they use similarly reliable processes. This core idea was maintained in the GBC.

The problem, as revealed in the preceding section, is that our knowledge of how the visual system works – especially of its neural components - is still so incomplete that it is impossible to construct an argument for the reliability of HP based on the reliability of the causal processes that contribute to it. It might be possible to make this argument at some point in the future once the many gaps in our knowledge of the human visual system have been filled in, but that doesn't help us with the project of trying to justify the reliability of (some) complex imaging technologies now. Does the (present) inability of the GBC to be used for this purpose mean that we are left with no way of establishing the epistemic reliability of technologies like PET and

CLSM? That it does not is clear when we reflect on the fact that the crucial part of the GBC was the need to identify reliability-making features of HP. This was what provided the required connection between a causal account of HP and an epistemic account and a justification of its epistemic status. However, similarity to the causal processes of HP is not a necessary condition for an instrument being reliable. If reliability is understood to be something like the avoidance of error in producing a representation of certain features of the world,[81] then what we need is to ensure that our instruments are capable of this. They need be similar to HP only in this property of being reliable.

The fact that not even HP itself can be justified by using a benchmark strategy makes it very clear that there must be other sorts of arguments that can be used to justify the reliability of our senses and our instruments. In the case of HP, its reliability can be justified by the fact that we have a great deal of knowledge about the conditions under which it is reliable and the conditions under which it is not.[82] We know, for instance, that there must be appropriate lighting conditions, that solid objects can hide things that lie distal to them, etc. In the case of uncertainty, we can also easily alter the viewing conditions by changing the light or moving closer, farther, or to a different side of an object. We can also use our other senses to be sure, for instance, that what we're seeing is a live dog rather than a stuffed one. We can also seek consensus from other people. In short, we can argue for the reliability of HP based on our enormous experience with it and our ability to determine when the appropriate working conditions are or are not present. While our experience with imaging technologies is vastly less than our experience with HP, we can still use most of these ways to establish their reliability.

[81] This is consistent with Hacking's map analogy (1983, 208) as well as with vision scientists' descriptions of the goal of HP: "The goal of perception is to account for systematic patterning of the retinal image, attributing features to their real world sources in objects and in the current viewing conditions." (Treisman and Kanwisher 1998, 218)

[82] See, for instance, Goldman (1986) and Alston (1993).

Our ability to manipulate the conditions may be somewhat more limited in these cases, but it is often not the case that conditions can be changed enough to adequately test the reliability of the system. As the last section showed, our knowledge of the sorts of physical and mathematical processes that are involved in these technologies gives them an advantage over HP when it comes to trying to establish reliability based on how the system works. Recall, for instance, the discussion in Chapter 2 about testing whether the application of attenuation correction algorithms to PET data improved the quality of the images. In this case both physical means (the use of a phantom constructed out of parts with known attenuation coefficients) and knowledge of the algorithms were used to establish the sorts of errors that result when attenuation correction is not performed. A detailed analysis of how reliability[83] can be established will form the basis of the next chapter.

[83] Questions of experimental validity will be distinguished from those of reliability. While the distinction has not mattered here, it will be important in what follows.

4. Reliability

4.1. Introduction

The last chapter argued that it is reliability independent of any sort of physical or causal similarity to human perception that can determine the epistemic goodness of any type of data collection process and so defines the scope of observation.[84] Thus, what is needed now is an account of what reliability is, and how it can be assessed. At the outset I want to make it clear that I am not aiming to lay out a general theory of reliability that can be applied to all possible domains and that can escape all possible objections.[85] What I will present is a pragmatic account of reliability that is: a) relative to the sort of discriminations that are needed for a specific purpose, and b) relative to the sorts of properties or features of the world that an instrument (including the human visual system) can get at. This pragmatic approach is motivated by the sorts of challenges presented to us by human perception and imaging technologies.[86] We use them to get information of a particular sort about the external world – about properties or features of the world such as the spatial location, size, motion, and color of objects[87] - and, usually, with a specific question or set of questions in mind. Answering a particular question requires that we get information about a specific set of properties with a certain degree of

[84] Recall that the previous chapter identified observation as any reliable means of data collection without requiring any causal similarity to human perception.

[85] I do not, in particular, make any claim that the account to be laid out here can defeat skeptical worries about, to take Colin McGinn's example, benevolent deities that would preserve our sensory input in the case that all material objects should cease to exist (1999, 8-9). Similarly, I am not worried about brains in vats, etc. In declining to engage with the skeptic, I share the view expressed by others who contend that these sorts of skeptical possibilities are not among the relevant alternatives that scientists are trying to discriminate between and so need not be addressed by an account of scientific evidence.

[86] This approach has a great deal in common with the pragmatic approach to laws taken by Mitchell (1997).

[87] Not all of these apply to every imaging technology, or to every application of a particular technology. We do not observe color using PET, for instance. I also think that the same account of reliability will also apply to other sorts of non-visual sorts of observation, though I will not explore that here. I have already denied that observation is limited to those processes that are directly available to the human senses and there is no clear way, on epistemic grounds alone, to discriminate between processes that are, for example, vision-like and hearing-like. I am happy to grant Hacking's claim that we can "see" with a microscope (1983), for instance.

accuracy and some minimal degree of resolution.[88] In order for us to use a particular instrument to make the sorts of discriminations needed to answer a particular question, there must be a match between the granularity of the world and the resolution of the instrument. Reliability, in short, requires both preservation[89] of the structure or features of the object and a match between the granularity of the world at which a particular question is directed and the granularity of the instrument.[90] The task of this chapter will be to examine what this sort of match consists of and how we can determine if the match is good enough for the data obtained using some instrument to be good evidence, useful in answering the question or questions at hand. The second part, assessing the reliability of an instrument for a particular purpose, often presents significant challenges in the case of imaging technologies[91] since, like unaided human perception, they are not perfectly reliable under all conditions[92] and the methods we have of assessing their reliability may be limited, particularly in cases where we have no independent access to the properties which we seek to represent reliably.

There are, of course, many existing philosophical accounts of reliability. While none is on its own able to resolve the problems that are set by PET and human perception, it will be helpful to review several of them that suggest desiderata for any account of reliability. Within epistemology, Nozick's (1981) counterfactual "truth-tracking" account of knowledge and

[88] Having higher resolution than is required does not prevent the question from being answered, though it may make it considerably more difficult. If I want to count the number of bacterial colonies on an agar plate, my unaided eyes will usually provide the most efficient means of arriving at an answer. A dissecting scope with ~10X magnification may help me more easily spot small colonies, but looking at the plate under 100X magnification will not provide any additional assistance and will slow the counting process down considerably.
[89] Features of the object need not be perfectly preserved, but must be preserved within some definable set of limits.
[90] The granularity of the world is more precisely stated as the granularity of the *description* of the world at which a particular question is directed. While the world offers some constraints on how we may divide it, and so on possible descriptions of it, there is no uniquely correct way of dividing up and describing the world (cf. Mitchell 2002).
[91] The next chapter will draw heavily on confocal microscopy since the issues to be addressed there (having to do with the concept of gaining visual access to phenomena and the sort of advantages it might confer) are particularly relevant to it. In the interest of space and non-redundancy I will restrict myself to PET in this chapter. There are both more and more complex difficulties with establishing reliability in the case of PET, so it will be more informative to use it as the case study here.
[92] Including the conditions under which they are commonly used.

Goldman's (1986) reliable process account of justification are among the most commonly referred to types of epistemic externalism. Both, however, have been subjected to serious criticism resulting in successive revisions to Goldman's account and a broadly held belief that Nozick's view is untenable.[93] While Goldman's account is the most successful reliabilist account, reliabilism in general has been accused of both having an inadequate concept of reliability and of lacking a principled way of identifying process types to which a reliability measure can be applied (usually known as the generality problem). Help with resolving these difficulties can be found by looking to the philosophy of science. Here, there is a considerable literature which deals with the reliability of evidence under the rubric of confirmation theory and philosophy of experiment.

There is considerable heterogeneity of approaches with key differences between those who: 1) promote a logical or a priori account of evidence (*e.g*. Carnap 1962, Hempel 1965, Glymour 1980) vs. an empirical account (e.g. Achinstein 1985, 2000, 2001; Woodward 2000; Mayo 2000; Bogen and Woodward 1988,1992), and 2) those who take reliability to refer to convergence on the correct (true or some surrogate notion) answer in the limit (*e.g*. Kelly 1996; Kelly, Schulte, and Juhl 1997; Harrell 2000)[94] vs. those who understand reliability to minimize the probability of error given a finite amount of data or time (*e.g.* Mayo 1996, 2000; Roush 2005).[95] Recent work in this area has argued strongly in favor of an empirical rather than a logical approach (Bogen and Woodward 1992; Woodward 2000; Mayo 2000; Achinstein 2000). Arguments to this effect point out that traditional accounts focus on the relationship between evidence and a

[93] Though Keith DeRose (1999) has incorporated some aspects of Nozick's account into his own contextualism and Sherri Roush (2005) advocates tracking accounts of both knowledge and evidence using conditional probabilities instead of counterfactuals.

[94] Some Bayesian approaches will also fall into this category, though convergence for them will be in terms of subjective probability such that convergence refers to the agent's certainty that her degree of belief will converge to 1 in the limit.

[95] These two axes of difference are clearly overlapping and are not intended to represent independent positions.

hypothesis rather than on the role played by the reliability with which the evidence is produced. Since I am concerned primarily with the latter question, I will focus on the second area of disagreement. As with reliabilist accounts within traditional epistemology none of the proposed accounts are without difficulties, but after considering in section 4.2 how each succeeds or fails to shed light on the problem of reliability of observational processes, I will argue in section 4.3 for an account that takes reliability to be objective, is based on limiting error rather than converging to the truth and distinguishes the issue of reliability from that of resolution.

With this account of reliability in hand, section 4.4 will assess how processes can be determined to be reliable in the specified sense. The key challenges for this task will be to deal with the generality problem, as mentioned above, and to respond to charges that current accounts of evidence are unable to accommodate PET (Bogen 2001, 2002). With respect to the first, I will claim that if we can specify what it means for an experimental set-up to be repeatable, we ought to be able to apply the same idea to epistemically relevant process types. With respect to the second, I will argue that by disentangling the question of resolution from reliability and by further distinguishing the instrument itself from the specific application, we are much better able to account for the epistemic value of PET.

4.2. Reliability

The intuitive sense of the term "reliable" is something like trustworthy or tending strongly to lead to correct conclusions. Reliable data ought to indicate to us that something is the case (is present or absent, has a particular value, etc.) when it is, and that it is not the case when it isn't. For data to be reliable, then, it must correspond to the phenomenon in some sort of regular way,

with respect to some specified set of properties.[96] A reliable process, in turn, is one that produces reliable data.[97] This notion of reliability can be identified in a great deal of writing, both philosophical and scientific. However, this general statement can be agreed to by proponents of considerably different proposals for how reliability and reliable processes should be understood, once fleshed out. I will begin by looking at how reliable process accounts in epistemology try to elaborate this notion. Next, I will examine some important accounts of reliability within philosophy of science. Finally, I will suggest how accounts from the two areas can complement each other and provide a starting point for the account of reliability that I will develop.

4.2.1. Internalism and externalism

Reliabilism in epistemology is usually taken to be the pre-eminent version of externalism. Externalist accounts are often taken to be ways of getting around Gettier-type counterexamples by assuring a connection between knowledge and truth. Several different reliabilist accounts have been proposed, but they all hold that an agent is justified in her belief that p if her belief that p was produced by a reliable process. Importantly, the agent need not know that her belief is produced by a reliable process in order to be justified in holding the belief. As such, reliabilism stands in contrast to internalist accounts of justification which require that the agent have access to and able to give reasons to justify her belief that p. Both internalism and externalism have been subjected to heavy and ongoing criticism and a complete defense of either approach would require solutions to some very significant problems – solutions for many of which I do not claim to have. However, the project of developing an account of reliability that is capable of

[96] Any particular type of instrument detects only certain objects or properties; it cannot produce reliable data about features of the world that it does not detect or (in some case) that are not causally related to what it detects.
[97] While there is an apparent symmetry in taking reliable processes to produce reliable data and reliable data to be those produced by reliable processes, it is obviously the process not the data that determines reliability. However, analysis of the data plays a crucial role in allowing us to assess the reliability of the processes that produce them.

supporting the epistemic status of observational processes does not require that the debate between internalists and externalists be solved.[98] My purpose here is not to answer the question of when we are justified in holding our beliefs, but instead to provide an account of the sorts of processes that are most conducive to correctness.[99] For this purpose an account that shares some, though certainly not all, features with the externalist project is needed. Before turning to Nozick's "truth-tracking" account of knowledge and Goldman's reliable process account of justification, however, I want to clarify what my account will and will not take from or contribute to the larger dispute between internalism and externalism.

Reliable processes are the most likely to produce knowledge if we use even a very coarse understanding of reliability as trustworthiness: even if a particular individual does not know that the process (e.g. PET) is reliable, there is still an objective fact of the matter about whether or not it is. The internalist and externalist disagree about whether or not an agent must know that the process that produced her belief is reliable, but both ought to admit that an unreliable process can't justify beliefs since it doesn't tend to produce true beliefs. (And if it does, it is by sheer luck.) So, like the externalist I need to have some way of characterizing what reliability means and what sorts of processes are reliable in this sense. In order to determine what processes are reliable (in whatever sense), one clearly needs to find and give reasons, and in the case of contested or competing knowledge claims, reasons may need to be given for why one method or instrument is more reliable than another. Thus, my account is not intended to defeat the internalist. After I have defended my account of reliability, it could certainly still be argued by

[98] One of the key goals of both internalist and externalist accounts is to fend off skepticism, but, as I said earlier, that will not be my concern here.

[99] I will use the term "correctness" to refer to the sought-after relationship between the data and the hypothesis or between the data and their representation of the phenomenon. I mean by this term, however, only that the relationship be one which is judged to be correct, whether one takes this to be truth, empirical adequacy, good predictive power, or some other notion.

the internalist that it is not the fact that my believing p was produced by a reliable process that provides the justification for my belief, but rather that the justification comes from facts that I can know by reflection alone.[100]

It is plausible to claim that our justification for some belief that is based on data obtained using PET or some other imaging technology is justified only if someone – though not every individual user of the technology – has internal access to facts about the reliability of the instrument for a certain purpose that justify this belief.[101] Such a claim would, however, need to allow that complete and sufficient facts about this reliability not be present in the mind of any single user but instead be distributed among workers who are themselves spread between multiple disciplines – statistics, computer science, physiology, etc. – and so possessed "internally" only by the scientific community more generally. This is not necessarily a problem for the internalist, however, as long as an argument can be made that justifies my reliance on pooled expert knowledge. Something like this can be found in what Brandom refers to as the "social articulation of the space of reasons" (1995).[102] Very briefly, Brandom insists that considerations of truth and reliability cannot be completely disengaged from the ability to give reasons, but that the giving of reasons ought not to be understood as a requirement to be met by each of us individually. We must recognize, first, that reliability is connected to the legitimacy of inference in that it concerns the reason for believing or knowing p and, second, that it is socially articulated inference. There are always two social perspectives involved: that of the one to whom knowledge (or standing in the space of reasons) is attributed, and that of the one

[100] Pryor (2001, 104) characterizes this basic internalist view as "simple internalism". There are many more stringent takes on internalism, but for our purpose the minimal view will be sufficient.
[101] If nobody had access to this sort of justification, no one would use the technology or at least not place much weight on the evidence gathered using it until some such justification is established.
[102] Brandom is not defending a strict internalist view and his emphasis on truth and reliability as external features of the world would have to be dealt with by any internalist wanting to adopt the more congenial (to them) parts of his account.

who attributes it. In some cases, I can attribute knowledge to another person because I can defend his reliability even though he cannot. The social distribution of reason giving ability, therefore, allows us to have knowledge in cases where we ourselves do not have the expert's reason-giving capacity.

Additional support for such an argument can be found in work on social epistemology (e.g. Shapley and Grofman 1984; Goldman 1999) that aims to show how certain modes of combining various sets of expert opinion can increase the accuracy of group belief. It might also be found by analogy to Putnam's causal account of reference according to which reference, after being fixed to samples of a particular natural kind at a dubbing, is to whatever has the same internal structure of the samples (e.g. having the chemical structure H_20 in the case of water). Those present at a dubbing are able to transmit the reference to others via communicative exchanges, these others can then lend the reference to yet other people, and so on. In this way, speakers who are ignorant about the internal properties of the kind in question can nevertheless use the natural kind term to refer to the members of the kind because underlying their uses are causal chains stretching back to a reference-fixing (Putnam 1975). Similarly, neither the oncologist who sends a patient to get a PET scan, the PET technician who performs the scan, the statistician who develops statistical techniques for modeling the data, the computer scientist who writes the software that gets used to process the data, the physicist who develops new radiopharmaceuticals for use in PET, nor the physiologist who develops tracer kinetic models of the movement of the radiopharmaceutical between various compartments (e.g. between capillaries and synapses in the case of a compound that binds serotonin receptors) individually has complete knowledge of the necessary and sufficient conditions for when and why a PET scan using a certain radiopharmaceutical and specific collection parameters and specific mathematical and statistical

processing methods will be able to serve as a reliable indicator of whether that patient has cancer. Each could contribute some of the needed conditions - and so, reasons to believe the scan was a reliable indicator of the presence (or absence) and location of cancerous lesions - but none could provide them all. However, this fact doesn't bear on the matter of whether or not PET is or is not a reliable indicator of cancer:[103] it is not the giving of reasons that makes the process conducive to knowledge but rather the reliability-making features of the process itself. Thus, what I need to provide an account of the reliability of evidence (or the reliability of the relation between data and phenomena) will have much in common with externalism, specifically with process reliabilism, but less in common with internalism.

4.2.2. Tracking the truth

An externalist account that has received a great deal of attention, though it has fallen largely out of favour in recent years, is Nozick's "truth-tracking" account of knowledge. According to Nozick's account of knowledge, belief counts as knowledge just in case it "tracks the truth", i.e. if it reliably covaries with reality across a certain range of close possible situations. After going through several iterations, Nozick presents the following analysis of knowledge:

> Let us define a technical locution, S knows, via method (or way of believing) M, that *p*:
>
> 1. *p* is true.
> 2. S believes, via method or way of coming to believe M, that *p*.

[103] As will be made clear later, the claim that PET is or is not a reliable indicator of cancer is far too broad. A better, though less succinct, way to phrase it is whether PET is a reliable detector of lesions of a particular type (e.g small cell lung carcinoma) and size in a particular location. The fact that knowledge is distributed among such a large set of disparate disciplines might conceivably pose a problem for the internalist if there is no effective communication between the groups. I will not consider this possibility, however, since I do not consider it to be a real problem. A related difficulty, however, is the use of PET scans in situations in which they are not reliable and taking the results to be reliable indicators of the presence or absence of some disease or condition. This is less of a problem in oncology, though the use of whole-body CT in otherwise healthy individuals as part of a supposedly thorough (and expensive) "check-up" has in recent years aroused concern. It is also a real concern in neuroimaging where PET images are used in legal courts and other public arenas as objective, supposedly reliable indicators of mental illness, personality disorders, etc. (c.f. Dumit, 2004).

3. If *p* weren't true and S were to use M to arrive at a belief whether (or not) *p*, then S wouldn't believe, via M, that *p*.
4. If *p* were true and S were to use M to arrive at a belief whether (or not) *p*, then S would believe, via M, that *p*. (Nozick 1981, 179)

It is the reliance on subjective conditionals that is perhaps the most serious problem with this account. McGinn (1999), for instance has claimed that since counterfactuals always have a dependent truth value, Nozick owes us an account of what makes the counterfactuals true. If, as McGinn claims, a satisfactory analysis must reveal the categorical facts upon which the counterfactuals depend, then the counterfactuals themselves are eliminable and the real work is done by the categorical statements. Nozick's strategy to deal with worries about the status of counterfactuals is to claim that their truth status depends on whether they hold in close possible world:

> "This point [about the power and intuitiveness of the subjunctive condition] is brought out especially clearly in recent 'possible-worlds' accounts of subjunctives: the subjunctive is true when (roughly) in all those worlds in which p holds true that are closest to the actual world, q also is true. (Examine those worlds in which p holds true closest to the actual world, and see if q holds true in all these.) Whether or not q is true in p worlds that are still farther away from the actual world is irrelevant to the truth of the subjunctive." (Nozick 81, p. 174)

Even if we put aside general worries about possible worlds, it is very difficult to see how this account could be applied to the cases we are interested in. Consider, for instance the subjunctive conditional, "If cortical area X were not involved in cognitive task Y, then this particular set of voxels would not have been active when subjects were asked to do Z." What are the closest possible worlds to ours in which cortical area X is involved in Y? And how are we to assess

whether or not the antecedent holds in all of those worlds? The standard account of closeness of possible worlds is in terms of the material facts and natural laws that hold in various worlds. So what we would want is to find worlds with the smallest possible difference in facts (made possible by a 'small miracle" or exception to a law in this one instance) and identical or, at most slightly different sets of laws and check the pattern of activation that would be derived from those facts using the laws in order to determine if (or at what distance of possible world) an area of activation changes. But notice how very strong the third and fourth condition are and the fineness with which we would have to specify the set of laws involved in producing a specific area of activation if we want to assess whether tracking holds for this PET experiment or for PET more generally. If we could derive the results (activation patterns) from the laws and material facts that hold in the close possible worlds, then certainly we must also be able to provide such a derivation for the actual world. But there is no way in which we can do this. As we shall see in section 4.4, that sort of fine-grained analysis based on derivation of results from laws (or from the particular algorithms that are applied to the raw count data) is simply not possible. We must supplement our theoretical understanding of the mathematics and statistics (as well as of the physical processes involved prior to detection of a photon) with experimental manipulations such as the use of phantoms in order to know what pattern of activation will be observed given a particular input. If we cannot predict exactly when (e.g. using what distribution of radiation density in the phantom or which reconstruction algorithm) or how the activation pattern will change in the real world, then we have no chance of assessing it in non-actual worlds with slightly different laws. Yet, it is my contention in this chapter that we can, at least sometimes, check to see if PET data is reliable. The details of how we can do this will be discussed later, but for now what matters is that Nozick's counterfactual analysis cannot possibly allow us to do this

– ever.[104] Thus, his account does not provide much assistance in judging the reliability of human perception and imaging technologies.

4.2.3. Reliabilist accounts

Goldman's reliable process account of justified belief is often taken to be the best developed version of reliabilism (e.g. Beebe 2004; Ginet 1985; Haack 1993; Harrell 2000), so I will rely on it to examine what reliabilist accounts might have to offer. In one of his early papers, he begins by saying that "reliability consists in the tendency of a process to produce beliefs that are true rather than false" (1979, 10). He goes on in that paper to provide several successive modifications of his definition of justification based on this general idea. Similarly, in work over the next two decades he continues to refine his definition of when a person is justified in his or her beliefs. Many of these refinements are motivated at least in part by consideration of skeptical arguments and so respond to difficulties that will not be addressed here. The crucial point for the project at hand is that throughout the successive refinements of his proposal, Goldman retains the conception of reliability as the tendency of a process to produce more true beliefs than false ones. I will only consider, therefore, one later version in which he defines justified belief as follows:

> S's belief in p is justified iff it is caused (or causally sustained) by a reliable cognitive process, or a history of reliable processes. (Goldman 1994, 309)

A cognitive process is reliable if it produces a sufficiently high ratio (>50%) of true to false beliefs. If a belief is produced by a process with a high truth ratio (i.e. by a process with a high

[104] That is, we could check *none* (vs. some) of our results at present. If we had much greater computational ability than we actually do – and much faster brains or computers – then it is possible that we could derive the needed results. However, since this is not currently the case and is not likely to be the case at any point in the foreseeable future, I do not take this to count in favor of Nozick's account.

degree of reliability), the belief has a high degree of justification. If a belief is produced by a process with a lower truth ratio, the belief has a lower degree of justification.

On the face of it, this seems to match very well with what I claimed above we would like to get from reliable data if we take the beliefs in question to refer to beliefs about the data. Data that has a high probability of correctly discriminating between relevant alternative claims about the phenomena will have a high truth ratio. As long as we have some means of assessing the truth ratios, presumably by reference to independent means of assessing the truth or falsity of our beliefs or data, we should be able to determine whether or not a particular process is reliable. It is in the details of how this should be done, however, that problems emerge.

Several objections have continued to be raised over the years to Goldman's account and have not been satisfactorily resolved with any of the revised versions of his proposal. One of these objections is the lack of precision of his truth ratios, in particular the fact that he has refused to specify a particular truth-ratio a process must have to be considered reliable. I am unconvinced that this is a significant problem, however. To begin with, while we would like reliability itself to be something that we can define in absolute terms even if few or even no actual processes are perfectly reliable; we would also like to be able to assess by how little or by how much a particular process falls short of this ideal. While producing false belief nearly half the time does perhaps run counter to most people's intuitions about what it means for a process to be reliable, it may be the case that certain processes do generate a relatively high proportion of false beliefs but that we also have the ability to identify which ones are more likely to be wrong and so increase the effective truth ratio. If we are able to isolate at least some of the false beliefs, in other words, we may be able to use the process in a more limited manner (i.e. not for the types of beliefs that have a higher chance of being wrong), but with a higher degree of justification. This, however,

raises the question of whether we are still referring to the same process or to a new, more tightly defined process that has a higher degree of reliability than the original process. Trying to resolve this issue leads to two more objections, both of which I take to be much more serious. The second objection is that the idea of the tendency of a process to produce a certain proportion of correct beliefs is vague. The third is what has come to be known as the "generality problem" and concerns the fact that reliability has to be ascribed to process types rather than tokens, but that there is no principled way of identifying a single relevant process type that is responsible for producing the belief for each process token. I will consider each of these in turn.

The problem of specifying what it means for a process to produce a certain ratio of true beliefs is that it seems like we must have a process that can be repeated many times in order to see how often it usually gets things right (i.e. enough times that we can perform a meaningful statistical analysis of its performance). This not only requires that we be able to specify the relevant process (bringing up the generality problem), but that we face up to the difficulty of what interpretation of probability to use. Goldman's reply to this challenge is that "general reliability is probably best understood as a propensity rather than a frequency. This avoids the possibility that actual uses of the process are numerically too limited or skewed to represent an intuitively appropriate ratio." (1986,49). Propensities are, of course, notoriously vague themselves as well as having a somewhat tenuous connection with what actually happens in the world. This, then, is a serious challenge. Similar difficulties involving the appropriate interpretation of probability will surface again later in the discussion of error statistics, so I will leave this problem for now and return to it at that point.

The generality problem arises, as I said above, because you cannot assign truth ratios to process tokens, but only to process types. A process token is, by definition, a unique event

(producing a single belief) occurring at a particular time and place; it makes no sense to ask if it would produce mostly true beliefs over repeated use.[105] Thus, in order to provide meaningful truth ratios, we need to refer to process types. Since each belief is produced by a process token, we must also be able to identify each process token as being an instance of a particular process type. This is the real source of the problem since for every process token, it seems that we can provide a seemingly endless number of process types to which it may belong. To illustrate this problem, Conee and Feldman (1998) offer the following example: suppose that Smith looks out his window, sees a maple tree, and forms the belief that there is a maple tree nearby. As long as the proper operating conditions for human visual perception are in place (i.e. Smith has normal eyesight, is not drunk or otherwise hallucinating, there is sufficient light, etc.), then it seems plausible that Smith's belief is justified. However, the token process that produced Smith's belief is a member of many process types, including but not limited to the following:

1. visual process
2. perceptual process
3. tree-identifying process
4. classifying tree by identification of leaf-shape process
5. process of retinal image with specific characteristics leading to belief that a maple tree is close by
6. process of identifying objects through a window
7. process of identifying trees through a window
8. process of identifying trees completely behind a solid obstruction

[105] If we were to do so, all processes would either have a truth ratio of 1 (if the token produced a true belief) or 0 (if the token produced a false belief). Goldman identifies this as the Single Case Problem and proposes that by using a propensity account of truth-ratio production it can be avoided.

And so on. If every process type to which a token could conceivably belong tended to produce the same truth ratio, then this might not be such a severe difficulty for reliabilism, but even cursory consideration of the above list shows that this is not so. Type (7), for instance would undoubtedly have a fairly high truth ratio (at least if we further specified that the trees were reasonably close), but type (8) will presumably have quite a low truth ratio since glass is a special kind of solid and most solid objects that completely occlude objects to be identified will cause a very significant decrease in the truth of identifications made using this process type. It is not obvious whether the actual process token was an instance of type (8), in which case his belief would seem to be unjustified, or of type (7) - or (1) through (6) – in which case it would seem to be justified. The heart of the generality problem, then, is that there seems to be no principled way to identify a single process type that is the epistemically relevant type of which a particular process token is an instance. Since the idea of a stable repeatable process is one that is also required in order to characterize experiment, it seems likely that in turning to accounts of reliability within philosophy of science, we may find some help in trying to resolve this objection as well.

4.2.4. Accounts of reliability within philosophy of science

In the above section, reliability was taken to be something that could apply equally to any belief-generating process. The same is true when we turn to philosophy of science. Science uses many different sorts of methods to try to get knowledge about the world and if reliability serves the ultimate goal of inquiry (whether taken to be correctness truth, empirical adequacy, accurate predictions, etc.) then it is something we ought to seek in all of our methods. Nevertheless, questions of what makes something good or reliable evidence – its relation to some phenomenon of interest – are often treated separately from those concerned with how evidence supports some hypothesis. We might take these to be questions about the reliability of observation and the

reliability of induction, respectively. I am primarily concerned here with the reliability of observation. This question is not entirely separable from that of inductive inference, but the emphasis is different.[106]

I will be concerned in this section primarily to examine the question of whether we ought to understand reliability in the context of human perception and imaging technologies as limiting logical reliability in the sense of Kelly (1996) or, instead, as finite probabilistic reliability (as advocated by classical statistics and by Mayo (1996, 2000)). Briefly, a method is characterized by logical limiting reliability if, for a given hypothesis, the method converges in the limit to the truth about that hypothesis on all possible sets of data that are consistent with the background assumptions. A method has finite probabilistic reliability if it generates erroneous results with a low enough probability given a finite amount of data.

On the face of it, it might seem that neither account is really what we want. Ideally, what we would like is a process that is guaranteed to get things exactly right[107] all the time. However, such processes are rarely, if ever, available to us. It is also the case, though, that in order for data to be useful – for them to allow us to identify relevant features of phenomena and discriminate between different hypotheses regarding the phenomena, we rarely need perfect accuracy and the highest degree of precision. But, if we acknowledge that practical considerations often make it

[106] As I indicated earlier, several authors have recently argued that an adequate account of evidence (usually observational evidence) must be empirical rather than logical (e.g Bogen and Woodward 1992; Woodward 2000; Mayo 2000; Achinstein 2000). For instance, Bogen and Woodward "deny that the use of data to correctly discriminate between competing phenomena claims requires that the data stand in some distinctive logical relationship to those claims of a sort that has been the subject of standard philosophical accounts of confirmation." (1992, 594-5). I agree that the types of processes by which we try to achieve correct discriminations of this sort are very heterogeneous and have more to do with the particular methods and circumstances we use than with any sort of logical framework. This is not to say, however, that logical frameworks are not also useful for other aspects of science, such as how different sorts of evidence (reliably obtained) can provide support for hypotheses or theories. However, since my interest here is with the reliability of evidence and of methods of obtaining evidence, I will not discuss here either hypothetico-deductive accounts or other strictly logical accounts such as Hempel's "satisfaction" theory (1965) and Glymour's bootstrapping (1980).

[107] With respect to the features that the process in question represents about the phenomenon.

impossible to actually obtain perfect data, how should we decide between different accounts of reliability? Should we prefer a method that is guaranteed to get things right eventually (even if we may not ever arrive at that point and, even if we do, won't know when we're there) or one that gets us acceptably close[108] to the correct answer, but may never get it exactly right? The answer will depend on what problem we're trying to solve. The problem that human perception, PET, and other imaging technologies set before us is whether or not the data they provide allows us to reliably discriminate between relevant alternative hypotheses about the things we are observing given a relatively small amount of data - a finite number of trials or a reasonably short viewing time (for visual perception). This is particularly the case for PET. PET is often used in clinical contexts where both the amount of data that can be collected and the computational resources that can be devoted to any one scan are severely limited.[109] The amount of data that can be collected is limited both by the fact that the amount of radioactive compound that can safely be injected into a patient is quite small and by the fact that the detectors themselves have upper limits on their count rates. Computational time is limited simply because of the fact that no one region and no one hospital or research institution has very many PET scanners. In order to fit in as many patients as can benefit from scans, time spent on the scan itself as well as on the processing of the data must usually be minimized as much as possible. In the end, then, what we want to know is whether the answer we arrive at after a relatively short series of trials is likely to be approximately correct. To address this question what we want is some sort of finite probabilistic reliability.[110] To see why, let us first look at logical limiting reliability to see why

[108] What is "acceptably close" will depend on the specific context. Some questions will require that more fine-grained discriminations be made and this may, in turn, require that the range of error be smaller.
[109] Similar constraints apply to research applications. In addition, cost often becomes a more significant factor when it is research funds rather than insurance that pays the cost of the scans.
[110] This term is borrowed from Steel (2005).

the sort of problem that it can help us with is not the problem that imaging technologies present us with.

Logical limiting reliability requires of a reliable method that it converge in the limit on the correct answer (again understood as truth, empirical adequacy, or some other criterion) in every context consistent with our background knowledge (c.f. Kelly 1996, Harrell 2000). It is itself a compromise of sorts since there are many ways to cash out "convergence", some of which might seem a better match with our intuitive ideal. We might have convergence in the following ways: by time t, with certainty, and in the limit.[111] Convergence by time t means that we will get the right answer by a particular time or deadline. This is the strongest criterion of success, but it is clearly too strong for most scientific inquiries: we usually don't know how much data or how much time will be required to determine the status of a hypothesis. Convergence with certainty is weaker, but still too strong for actual practice. According to this standard, a method should output a specific mark (Kelly uses "!") right before it outputs the correct answer. We don't know how long it will take, but we'll know once we get there. Unfortunately, our methods don't usually let us know when we've reached the end, with a "!" or otherwise. Weaker still is the idea of convergence in the limit. A method converges in the limit if there exists some time after which it will forever output that particular answer. We cannot be certain about when this point has been reached however, since any finite sequence of outputs is consistent with any answer in the limit and, of course, we will never actually reach the end of an infinite data stream. Kelly claims that many scientific hypotheses fall into this category and that it is useful to be able to determine which hypotheses will allow knowledge in the limit even if we're not sure that we've got it yet.

[111] Kelly also considers gradual convergence, but this notion is weaker than the one he chooses, so I will leave it aside. There are also 3 different ways to cash out "the truth": verification, refutation or decision.

I do not deny that convergence in the limit may be a theoretical possibility for many methods including imaging technologies, though the constraints of our actual scientific practice will often restrict us to a relatively short data stream. My contention that logical limiting reliability is not the sense we need to assess the reliability of imaging technologies stems from the fact that this sense of reliability is designed to address questions about underdetermination rather than about the problem that imaging technologies confront us with. Kelly is concerned primarily with the problem of induction and, more specifically, with identifying the difference between theories that are underdetermined and those that are not. He proposes that underdetermination is exactly the impossibility of logical reliability (Kelly 1996, 30). The framework he goes on to develop, then, is intended to examine the question of whether there are methods that can be shown to be logically reliable. If no such methods exist, then we are faced with underdetermination, but if some methods are reliable in his sense, then at least some knowledge can be freed from the specter of underdetermination.[112] While it may be very interesting to explore the logical structure of problems and methods that will converge to the correct answer in the limit, this understanding of reliability is of little use for the question of whether or not some method is likely to get the correct answer (to whatever degree of approximation is acceptable for the purpose at hand) after a short series of trials. Any initial data stream is consistent with any answer in the limit, so we have no reason to think that the finite amount of data we have available to us stands in any particular relationship to the correct answer. We would like to be able to claim that individual uses of visual perception or some imaging technology are reliable, not just that they are methods that are reliable in the limit. Limiting reliability, therefore, is not useful for evaluating the reliability of currently used methods of observation.

[112] This is particularly interesting at the highest level of abstraction where the methods in question are very general inductive methods or what he refers to as "complete architectures for induction" (Kelly 1996, 36).

The other way that we might choose to understand reliability is as finite probabilistic reliability.[113] To meet this standard, a method must produce erroneous data with an acceptably low frequency given a finite specified sample size. Defenders of orthodox statistical methods (e.g. Mayo 1996) claim that this standard is the one that should be met in order to be consistent with actual scientific practice. Does this variety of reliability do any better for the problem of getting the (approximately) correct answer with a limited amount of data? I think it does, but we need to be careful about how we describe it. Above, I said that finite probabilistic reliability requires of a reliable method that it generate erroneous results with an acceptably low rate given a specified and finite sample size. The key point is that we want to know the chance of error. But orthodox statistics doesn't tell us how likely we are to be wrong. This is a common misconception of what p-values[114] tell us (Gigerenzer 2000; Mayo 1996), but in fact they tell us only the probability that we would get the data we did if the null hypothesis were true. In other words, if a PET study identifies a particular region of the brain as involved in a particular task with $p<0.01$, it doesn't tell us that there is less that a 1% chance that this area isn't involved in the task, but only that we can would have less than a 1% chance of getting this data if the area weren't involved in the task. *If* we want the chance that we're wrong, then what we need is to incorporate the base rate into the calculation (to get the positive predictive value of some result).

Mayo (1996) has argued vehemently that it is not what we want. Instead, she claims that what we want is for error statistical analysis to assess whether or not the evidence is effective at

[113] There is a third option: limiting probabilistic reliability. According to probabilistic limiting reliability, a method must converge on the right answer with probability 1 (e.g. Spirtes, Glymour, Scheines 2000). In the case of Bayesian convergence theorems where probability refers to the agent's degrees of belief, this means only that the agent must believe that the method will converge, not that it actually will (Steel 2005; Kelly, Schulte, and Juhl 1997). If a method can meet the former standard, it will also meet the latter, so I will only discuss logical limiting reliability here.

[114] The p-value (probability value) of a statistical hypothesis test is the probability of getting a value of the test statistic of equal or greater magnitude than that observed (by chance) if the null hypothesis were actually true.

ruling out the error in assessing the particular hypothesis under test. But base rates can clearly be relevant to the question of how likely we are to be making an error. It is not entirely clear, then, why she denies that the positive predictive value of a result is something that we ought to consider in characterizing the error characteristics of some method. In cases such as the interpretation of medical diagnostic tests, it certainly seems as though we ought to incorporate the base rate if this information is available to us (though in many cases it might not be). If, for instance, we know that the sensitivity of HIV testing is 99.8 % (the test will be positive in 998 out of 1000 people who are infected with HIV), that the rate of false positives is 0.01% (or 1 in 10,000, and that the prevalence of HIV in low risk populations is also 0.01%, it would seem to make a relevant difference to the error status of the test in the case of an individual who tests positive if we know that that person is in a low risk or high risk group. While the positive predictive value of a positive test for someone in a high risk group is 99.8% (or is at least near that, depending on how high risk the person's behavior is), for a person in a low risk group is only 50%.[115] In other words, we are very likely to be wrong in the latter case, but only minimally likely to be wrong in the former.

Prevalence of HIV in lower-risk population	# of true positives per 10,000 tests	Positive predictive value of a positive test
0.01%	1/10,000	50%
0.02%	2/10,000	67%
0.05%	5/10,000	83%
0.1%	10/10,000	91%
1.0%	100/10,000	99%

Table 4.1 Effect of base rate.

[115] This example is taken from Gigerenzer (2000, 81).

Moreover, failure to take the base rate into consideration can lead to very significant error even with relatively small differences in the base rate. Table 4.1 shows the effect of relatively small differences in the prevalence of HIV in a population on the positive predictive value of an HIV test (all other values remaining unchanged from the scenario described above). It seems to me to be unproblematic to admit base rates into calculations such as these. It is not even necessary to admit subjective probabilities in order to do this, so to admit this point would not require that Mayo give any ground on this matter (the point against which a great deal of her criticism of Bayesians is directed). While I agree with Mayo that probability and reliability ought to be objective, the prevalence of certain conditions (such as HIV infection) among various classes can be objectively measured as can an individual's likely membership in a particular class (i.e. by a survey of their patterns of behavior).

In addition to her dismissal of the relevance of base rates, another difficulty with Mayo's account as it stands is that it is not entirely clear how she understands the idea of probability. She does say that she favors a frequentist interpretation but is silent about whether she means by this a limiting relative frequency or a finite relative frequency approach. If she means the former, then this version of reliability will end up collapsing into a version of logical limiting reliability and so will have little use for imaging technologies. If she means the latter, then she owes us a solution to difficulties with this interpretation such as the Gambler's Fallacy according to which if a gambler accepts a finite frequency account according to which the statement that "the probability of heads is ½" means that in a reasonably large, but finite sequence of coin tosses, the frequency of heads will be very close to ½, then if he tosses a coin he believes to be fair 1000 times and gets all tails, he will bet on heads for the next 1000 tosses. Mayo offers no such account and, indeed, seems to avoid specifying how she interprets relative frequencies.

This is perhaps not surprising, since the finite frequency account generally finds few defenders. However, in a recent paper, Glymour (2003) has provided a defense of this view. Following a discussion of the merits of data analysis in terms of uncertain but bounded error (which he claims likely lost out to least squares analysis due to the latter's greater computational tractability), Glymour contends that:

> "the finite frequency story is something else besides a definition of "probability", that it is a compressed account of how inferences from data may be made with the aid of the mathematics of probability, but without the obscure thing itself" (2003, 249)

and that we ought to take this interpretation:

> "as a proposal to use the language and mathematics of probability to approximately describe actual or potential finite populations, and as a means of generating definite, nonprobabilistic hypotheses" (2003, 249)

Glymour's way of dealing with the Gambler's Fallacy is to say that the reasonable gambler must either reject the distribution assumption that, for a reasonably large sample in which the normal distribution is not approximated, a larger sample will still be normally distributed, or, if that is impossible (e.g. in the case of a population from which successive samples are taken and not replaced), make the assumption that the next 1000 tosses will be heads with perfect rationality. In other words, after acquiring a significant amount of data in which the expected frequency of different possible outcomes fails to be obtained, rather than continue to believe that each successive trial has the same chance of a particular outcome (e.g. that the probability of heads on each subsequent trial will be ½) it is more reasonable to doubt that expected pattern of results (the distribution) was correct to begin with. Or, if it is certain that the distribution is as originally hypothesized, the reasonable person will acknowledge that the unlikely pattern of

results obtained up to that point requires that a similarly skewed series of outcomes must follow. In neither case would the gambler be making an illegitimate assumption that the chance of obtaining a particular outcome on the next trial or set of trials depends on the outcome of prior trials.

The most interesting part of the proposal, however, is how it allows us to interpret statements about the degree of error or approximation in large but finite samples: they are not claims about probabilities but rather about uncertain but bounded errors in some feature of the empirical distribution of an actual or potential finite frequency distribution. The more parameters (with their respective errors) contribute to the calculation of this feature, the larger the bounds of the error will tend to be since the uncertainty in measurement of each will propagate (additively or multiplicatively depending on their relationship to the quantity being calculated) through the calculation. Thus in the gas law example Glymour discusses (2003, 242-3), if we want to calculate the pressure of a sample of a gas at time t_2 when we are able to measure the temperature and volume of the gas at t_2 as well as the pressure, volume, and temperature at some other time t_1, we simply calculate:[116]

(1) $$P_2 = (P_1V_1T_2)/(T_1V_2)$$

If each quantity has a value of 100 and an error bound of 1, we can calculate with certainty that $95.118<P_2<105.122$. The degree of uncertainty, in other words, has expanded from 1% in the individual measurements, to 10% in the calculated value for P_2. If we were then to measure P_2 and found it to lie outside the bounds of error, say 106.5, then we would have to either reject the gas law or, alternatively, the assumption that the gas sample was the same at the two time points.

[116] The ideal gas law states that $PV=KT$ where K is a constant, P is pressure, V is volume, and T is temperature. K is constant for any species of gas, so for any one sample $K=P_1V_1/T_1=P_2V_2/T_2$. Upon rearrangement, this gives equation (1).

The same idea can be applied to the measurements and calculations involved in generating PET data. If we are able to calculate the error bounds for the measured quantities and know the calculations that are carried out using these quantities to produce areas of activation in the final image, then we can set error bounds for these areas of activation with respect to their location, intensity (represented by color), and time of appearance and/or disappearance (when relevant). Calculation of the error bounds in the final image requires both knowledge of the nature of the calculations, and measurement of the error for input measured quantities. Of course, the computations involved are far more complex for PET than for the gas law example and there is no simple and general way to determine the error bounds for PET. Because different algorithms and different settings can be used at various points in data collection and processing, the error bounds must be calculated independently for each. To get a better understanding of how this can be done, let us look at a comparison of spatial normalization techniques for PET.

An important part of the resolution of PET for studies of cognitive function has to do with the mapping of individual data sets onto standardized reference spaces. Many such spaces are probabilistic atlases that combine the anatomical features of multiple subjects in order to construct a brain map that will allow investigators to calculate the probability that a specific point of interest (in an MRI or PET data set, for instance) is within a particular anatomical structure. Glymour's version of the finite frequency interpretation allows us to understand these probabilistic brain atlases as containing definite, empirical claims about the location of particular anatomical features. When an investigator takes a particular point (area of activation) in a PET data set from an individual subject, maps it onto a standardized atlas, and finds that it falls outside the error bounds for the location of a particular structure, she can reject the hypothesis that the task that produced the activation involved that particular structure.

While this sounds straightforward, the process (spatial normalization) of getting to the point of identifying a particular area of activation as falling outside the error bounds of a structure are far from simple and are not easy to characterize in terms of their own error bounds. It is widely recognized that there are considerable anatomical and functional differences between individuals. Anatomical differences are present at the macroscopic level – the size and shape of the entire brain as well as of specific regions are variable – as well as the microscopic level of cellular architecture. Functional variation is superimposed on this structural variation since different areas may be used to perform the same task in different individuals. Relationships between structure and function (or the lack thereof) cannot be investigated without reducing as far as possible the structural variability between individual subjects. Accordingly, virtually all studies in cognitive neuroscience require a spatial normalization step in which data acquired in different individuals are mapped onto a common neuroanatomical reference space.[117] The original Talairach transformation (Fox et al., 1985) is restricted to linear transformations, but more recently several nonlinear brain warping procedures have been developed to match a given brain volume onto a standardized one. These procedures are based on different mathematical techniques and can be divided broadly into intensity-driven and model-driven approaches according to the sorts of features that are used to map one brain onto the other (Toga 1999).

[117] The availability of a good reference space is also an important issue. While many investigators still use the Talairach-Tournoux atlas, it has been extensively criticized for being based on one brain – that of a 60-year old female. Other references now in use include the MRI atlas of the Montreal Neurological Institute and the Human Brain Atlas. A new probabilistic map is currently being produced by Zilles and colleagues that will incorporate not just post-mortem MRI data for 15 brains, but also microstructural information generated from histological analysis (see Abbott 2003). After the MRI scan, the brains are embedded in paraffin, cut into 20μm thick sections (of which there are 5000-8000 per brain), and every fifteenth section stained to visualize the cell bodies. Every sixtieth section (i.e. one section every 1.2 mm) is imaged, its contours are morphed back to those of the *in situ* brain (since sectioning tends to distort the shape of the sample in just the same way as does cutting through any relatively soft object even with a very sharp knife), then the number and distribution of cell bodies are counted and a computer used to identify borders of distinct anatomical areas by searching for sudden and statistically significant changes in the number and distribution of cell bodies in corresponding coordinates of sequential sections.

Intensity-driven procedures define some measure of similarity[118] between the specific individual brain and the reference or target brain, and then adjust the parameters of the deformation until the value of the chosen measure is maximized. Model-driven approaches, on the other hand, start by building explicit geometric models that represent specific anatomical elements[119] in each of the brains to be warped, then parameterize each element and use these to guide the transformation of one brain volume to the other.

The precise details of these procedures are not important here: what does matter is that there is no way to calculate the error for each variable that enters into the procedures and, accordingly, no way to straightforwardly define the error bounds for the final spatial distribution calculated with a particular procedure as we were able to do for Glymour's gas law example. What we really want to know is how close a procedure can get to performing a perfect brain match as judged by residual anatomic variability (size of the bounded error) between subjects after spatial normalization. However, there are relatively few studies comparing different spatial normalization procedures and methods for measuring error in the final map.

Crivelli et al. (2002) compared the performance of four common normalization procedures for warping individual MRI brain volumes onto a standard reference template (the Human Brain Atlas). The merit of each spatial normalization procedure was assessed by using tissue segmentation as the criterion of success. Each method was used to identify[120] each voxel in MRI data sets as grey matter, white matter, or cerebrospinal fluid (taking the tissue classification performed on the Human Brain Atlas template as the gold standard). They then quantified the

[118] Some of the similarity measures that have been used are normalized cross-correlation (Bajcsy and Kovacic, 1989; Collins et al., 1995, 1995), squared differences in pixel intensities (Christensen et al. 1997; Woods et al., 1998; Ashburner and Friston 1999), and mutual information metrics (Kim et al., 1997).

[119] These include functionally important surfaces (e.g. Szeliski and Lavallee, 1993; Thompson and Toga, 1996; Davatzikos 1996) , curves (e.g. Monga and Benayoun 1995; Subsol, 1999), and point landmarks (Bookstein, 1989; Amit et al., 1991).

[120] Each voxel is assigned a probability of belonging to a particular tissue class then a threshold (in this case, 50%) is set for classifying each voxel as a unique tissue type.

degree of spatial overlap between the template and each MRI volume for each tissue class and for each procedure. In addition, the impact of different spatial normalization procedures on functional maps was investigated by taking PET data sets for the same individuals for whom the spatial normalization and tissue segmentation had been performed (using MRI data) and looking at the overlap in the volumes judged to be active according to the four different procedures. The results showed that there are differences in tissue segmentation between the spatial normalization procedures, but that the consequences of these differences were much greater when high resolution functional maps (FWHM ~6 mm) are used than at lower resolution (FWHM ~10 mm). Whereas 42.8% of the total activation volume was shared between the four methods for the low resolution functional maps, only 6.2% was shared at high resolution. No differences in the number of activated areas were observed, but the location of the active areas was significantly different between the different methods, creating problems for trying to distinguish precise activation areas within the same anatomical area.[121]

This means that unless there are good reasons to believe that one normalization method is better than another for a specific experimental question, a probabilistic map generated using one particular procedure may not correctly identify the relevant error bounds. If a map using multiple methods is used, however, the error bounds will be very large, even at low resolution. Which strategy is preferable will itself depend largely on the question. If, for instance, we were interested in whether or not different classes of subjects used widely separated brain areas to perform the same task, the wide error bounds might still provide sufficient resolution to answer the question. If the two classes showed non-overlapping areas of activation (including the error bounds), they would have been shown to involve different areas. If, however, the question was

[121] They do not claim that the procedure judged to be best for tissue segmentation is always the best choice for any experimental question.

more specific – about, for instance, the differential involvement of precise areas whose error bounds overlapped using either a single method or a compiled map – then the evidence would not be able to be used to reliably discriminate between the hypothesis that the same area is involved in both classes and the hypothesis that different areas are involved.

4.3. What characteristics must reliability have?

To summarize the position at which we have now arrived, an account of reliability that will help to solve the challenges presented by human perception and imaging technologies must have the following characteristics:

- It must be an objective relationship between the data and the features of the world it represents
- It cannot be characterized in strictly causal terms since we need to be able to describe complex instruments (including the human visual system) that involve both causal processes such as interactions between light and physical objects and statistical or mathematical processing steps that are not causal in any clear sense.
- It must recognize the fact that reliability is usually not an end in itself, but is instead a requirement for the achievement of another goal: discriminating between different (relevant) possibilities.
- It must provide us with the ability to make these discriminations based on finite amounts of data with a low frequency of error. In assessing the likelihood of error, probabilities should be understood as finite relative frequencies.

These characteristics, however, are not sufficient. In particular, more needs to be said about what is required for data to allow us to discriminate between different hypotheses. It is in allowing particular sorts of discriminations to be made that data or an instrument are reliable for

a particular purpose. I propose that an account of the reliability of imaging technologies requires three parts:

1. the concept of resolution must be distinguished from that of reliability
2. specification of the purpose-relativity of reliability
3. understanding reliability as a relation between the granularity of the world required to answer a particular question and the resolution of the data or instrument.

4.3.1. Resolution and purpose-relativity

An account of reliability must make a distinction between reliability and resolution and be careful not to subsume the latter to the former. Resolution is the smallest interval (spatial or temporal) at which two points can be distinguished by a particular instrument or process. In terms of spatial resolution, the increased magnification that we get from using a magnifying glass or microscope doesn't on its own allow us to see *more* than we can with our naked eye; we just see the same thing bigger (and less of it within a single field of view). Think, for instance, of zooming in on a fairly low resolution digital image on your computer. The more you zoom in, the more the image gets pixilated. You see an image made up of bigger rectangles, but each rectangle still only has a constant greyscale or color intensity, you don't see more detail in the image.[122]

In order for increased magnification to really be useful, we also need a corresponding increase in resolution so that we can distinguish finer detail in the object under investigation.[123]

[122] This simple story is true only for the specified low-resolution image. If you start with a high resolution image, this will still happen eventually, but if the resolution of the original image was beyond that of the unaided eye, you will be able to discern more detail as you first begin to zoom in.

[123] Increasing image size beyond the resolution of an instrument is often referred to "empty magnification" and confers no benefit on the observer.

It may not seem obvious that reliability is independent of resolution: someone might well object that if one instrument allows me to distinguish finer-grained spatial or temporal features of some phenomenon than another does, then surely it is more reliable. But to see that an increase in resolution alone makes no difference to reliability, consider the following. Suppose that I am interested in studying the bacterial population in a local pond. I am especially interested in knowing whether the proportion of a particular pathogenic species increases under particular environmental conditions. To do this, I will take samples of pond water at various times, put drops of the water on slides, and look at them in a standard light microscope.[124] Suppose, that the pathogen that I'm interested in is a rod-shaped bacteria while all the other pond species are round so that as long as I can visually discriminate between a rod and a spherical bacterium, I can distinguish the pathogen from the other bacterial species. Setting aside questions of sampling technique and other statistical questions, let us focus simply on the reliability of my visual identification of the pathogen. I cannot see any of the bacteria with my naked eye, but suppose that they are quite large bacteria so that I can easily tell the difference between a rod and a sphere at 100X magnification. All I see at this magnification is a smooth outline of a rod or a cone, but that is all I need to discriminate between the pathogen and all non-pathogenic species. Now suppose that, out of curiosity, I switch to the high power objective which gives me both an increase in magnification and in resolution (i.e. the high powered objective lens also has a higher numerical aperture than the low powered objective lens). Now I can see fewer bacteria in each field, but the rods no longer appear to have the completely smooth surface that they seemed to have under low power but are instead a bit uneven and rough-looking.

[124] Obviously this example is grossly oversimplified. Visual discrimination of different species would never be the sole mode of species identification and would likely be performed only after some sort of staining process (e,g, Gram staining).

What has happened as I moved from my naked eye to looking at the bacteria under low power to looking at them under high power? I have increased the magnification and the resolution in each step, but have I increased the reliability of the observational method? To even consider this question would make no sense on a anthropocentric empiricist account since we would seem to be obligated instead to defend the claim that we haven't *lost* reliability in moving along the chain away from unaided human perception.[125] However, since the previous chapter showed that the reliability of an instrument does not depend on its bearing some physical or causal similarity to human perception, the question should now seem entirely legitimate. My naked eye sees just some water - maybe clear, maybe cloudy, maybe with some algae or debris in it, but for all I know there are no bacteria of any kind in it. Using the low power objective of the microscope, I can easily distinguish the rod-shaped pathogens from all other bacteria. So as not to prejudice the example, let's further specify at this point that I occasionally misclassify a rod as a sphere, so that I am not perfectly reliable (though this need not indicate imperfect reliability of the microscope itself). Does this mean that my eyes are less reliable than the microscope on low power? This would certainly be a bad result for the empiricist.[126] However, we are not justified in coming to that conclusion, so the empiricist can rest easy on this count at least. What the example shows is that the human visual system is unable to make certain sorts of discriminations that can be made with a fairly high degree of reliability using the microscope. This is not due to a failure of reliability of human visual perception, but due to its insufficient resolution.

[125] An alternative empiricist strategy at this point might be to claim that only the sorts of discriminations that can be made using human senses are reliable. I see no way that this sort of distinction could be made in any principled manner, however.

[126] Since this would imply that, other factors affecting reliability being equal, the better the resolution of an instrument the better the reliability. So not only would the light microscope be more reliable than human perception, but the electron microscope would be more reliable than the light microscope, and so on. Since, as a very general rule, the higher the resolution an instrument has, the less similarity of any sort it bears to human perception, this would indeed be very bad for the empiricist.

We might be misled by stating the problem in the following way: unaided human perception is not a reliable instrument for detecting rod-shaped pathogens. This statement is true, but it is true only because reliability is to be characterized with respect to a particular purpose. The reliability of an instrument must be understood to be connected its ability to make the kinds of discriminations a particular purpose requires. If we try to require that perfect reliability discriminate between all possible states of affairs, then the only instrument that would stand a hope of coming anywhere near this would be one that can detect all possible properties (not just visually accessible features such as size, color, shape, and motion, but mass, gravity, chemical composition, etc.). This would preclude any human sense from being even remotely reliable and that just seems wrong. While I claimed in the previous chapter that it is important to recognize that human perception can fail and has something like proper operating conditions, I also claimed that it is usually very reliable. To do otherwise simply flies in the face of our vast experience with it. While an instrument is reliable for a specific purpose, the concept of reliability is still an objective relation as long as we keep in mind that a particular instrument can detect only certain properties. Each takes a specific input and produces a specific output which may but need not represent all of the properties possessed by the input. Reliability does not require detecting every property. But along the same lines, an instrument need not have finitely small resolution in order to be maximally reliable. To see this point, let's return to the microscope example.

In switching from low to high power, the rods that had seemed to have a smooth surface turned out to be rough and uneven (though still rod-shaped overall). Being able to discern this feature of the bacteria didn't allow for better accuracy in distinguishing between pathogens and non-pathogens, so it was not more reliable in terms of allowing that discrimination. However,

one might want to claim that using the high power objective is (objectively) more reliable in that it allows us to discern more fine-grained features of the bacteria, even if this additional information is not required for the discrimination we want to make. If this objection is to succeed, however, it must be because the more detailed surface is a better indicator of some feature than the less-detailed surface. But why should the less detailed surface be less reliable? Essentially what the smooth surface represents is the average amount of surface disruption over a stretch of bacterial cell membrane corresponding to the resolution of the instrument used to detect it. The only way that an average over some area is less reliable than the set of individual values is if the averaging procedure itself introduces error. As long as it does not, then lower resolution does not imply lesser reliability.

That said, however, the fact that reliability is assessed relative to a specific purpose means that there may well be purposes for which switching to the high power objective does increase reliability. If, for instance, there were actually two types of rods in the sample, one of which was still smooth and the other rough under higher magnification, we would be able to make discriminations that we could not have made under lower power. In general, it will be the case that higher resolution will increase the number and type of discriminations that can be made. Properties of objects that could not be distinguished with lower resolution can now be discovered. However, if the resolution of the instrument is greater than the granularity of the representation of the world at which discriminations must be made in order to answer a particular question, the increased resolution might actually slow or even prevent making the required discriminations. Too much fine-grained detail can obscure the relevant similarities or differences at a coarser grain.

4.3.2. Granularity match vs. mapping

Granularity is a characteristic of representations. We can refer to both the granularity of the representation of the world that a particular question is directed at and to the granularity of the representation of the world (data) that an instrument generates. For the sake of ease, I will refer to these as the granularity of the world and the granularity of the instrument respectively. The granularity of a representation is the smallest object or unit required to address the question of interest.[127] The larger the spatial and temporal scale of the aspects of the world which must be distinguished in order to answer a particular question, the coarser the grain. A question about the effect of annual fluctuation in berry crops on the size of grizzly bear territories, for instance, might not require that one investigate anything smaller than individual organisms (bear and berry), while a study of herpes egress from cells would require that questions be addressed towards and data obtained about sub-cellular entities and events. The granularity of an instrument *matches* or is sufficient for a question if it is capable of providing evidence about the smallest objects needed to address that question.

Normally, though, we speak of the resolution of an instrument, not its granularity. What is the relationship between the two? Often the two will coincide, but this is not always the case. The granularity of the instrument is related but not always identical to its resolution. In some cases, the granularity with respect to certain features may be higher than the resolution; *i.e.* some kinds of questions about objects that are smaller than the resolution of the instrument can be addressed. This occurs when these objects are coordinated in some way so as to allow detection of certain of their features despite the objects as a whole being too small to individuate. For instance, pictures taken of the stands of a football stadium from a blimp high above the ground could be used to make a reasonable estimate of the proportion of fans supporting each team

[127] In its question relativity, my account of granularity and reliability resembles the approach to laws taken by Mitchell (2000).

(assuming that all the fans wore the colors of their team). No individual fan could be distinguished, but patches of color representing groups of like-dressed individuals could be seen and the overall amount of the stands filled with each color calculated.

A similar situation can occur with biological imaging technologies. The spatial resolution of a PET or fMRI image is the voxel size.[128] It is limited by many factors including intrinsic characteristics of the receptor (e.g. for PET, the type of scintillating crystal and the size and geometry of the detectors), selection of the reconstruction algorithm, and spatial blurring caused by both motion of the subject and by biological or physical features of the system upstream of the detectors (e.g. the distance traveled by a positron in tissue before it annihilates).[129] Resolution here refers simply to the fact that voxels are the minimal spatially discriminable unit since a separate numerical activity value is calculated for each voxel.[130] This number represents the average activity for the spatial area[131] of the object – the cube of brain or other tissue – corresponding to that voxel.

The granularity of the instrument refers instead to the size of the units of data (i.e. the number of spatiotemporally discriminable units or voxels) that can completely represent some quantity.[132] Sometimes the granularity of PET (or fMRI) may be equal to the resolution, but often it is not. Some quantities such as blood flow cannot be represented in a single voxel so the granularity will span many pixels. In other cases, features of the PET detection system itself

[128] Voxel size does not refer to the dimensions taken up in the image by a particular voxel, but to the volume of the object (e.g. 2 mm^3) that each voxel represents as a single numerical value or colored square. The size of the image itself (and, accordingly, the size of each voxel in the image) can be made smaller or larger without changing the voxel size in this sense.
[129] Sanchez-Crespo, Andreo, and Larsson (2004).
[130] Notice that the resolution is not specific to the image – it applies equally to the data presented as a set of numerical values for each voxel.
[131] Each voxel also has a temporal dimension, reflecting counts obtained over some period of time. Temporal aspects of imaging will be discussed more in the next chapter.
[132] The granularity of an object or event will always be relative to some quantity such as neural activity in response to a particular stimulus: its granularity is the smallest area that completely contains that quantity.

mean that only structures that span two or more pixels can be adequately measured. The idea that only objects larger than twice the resolution can be properly investigated is captured in the specification of "full width at half maximum" (FWHM) that accompanies most PET data. Structures smaller than this have not generally been believed to be able to be interpreted reliably since they are significantly affected by the surrounding areas either through partial volume effects or averaging out of very small areas of high activity within a single voxel.[133] However, recent work has shown that this might not always be the case and that information about structures such as orientation columns that are smaller than a voxel can be identified using fMRI data[134] by using multivariate pattern recognition to identify patterns of activity that occur across space, from multiple voxels (e.g. Kamitani and Tong 2005; Haynes and Rees 2005; Cox and Savoy 2003).

There has been considerable recent interest is in the mapping of putative orientation columns in humans. In the visual cortex of non-human mammals, neurons with similar response properties to lines with a certain orientation (i.e. that are oriented at a particular angle within the visual field) have been shown to be clustered into columns These are referred to as orientation columns and are about 300-500 μm in width.[135] Questions about individual columns are questions referring to a granularity of the world of about half a millimeter. It is expected that questions about orientation columns in humans refer to or require a similar granularity. But mapping such structures in humans has proven to be difficult since the invasive techniques that have been used in cats and monkeys cannot be used in humans and non-invasive neuroimaging

[133] Partial volume effects are due to the fact that the count number for any given voxel is subject to interference or spillover into that voxel from adjacent voxels that have significantly higher activity. They result in some voxels being represented as having more activity than they actually do. Areas of high activity that are smaller than a voxel, on the other hand, will get missed since the activity of a voxel reflects the average activity over the whole of the area represented.
[134] These recent studies have all focused on fMRI, but the same sort of techniques could also be applied to PET.
[135] Vanduffel et al., 2002.

methods (primarily fMRI in this case) have a spatial resolution at the level of millimeters, not hundreds of micrometers.

However, one lab has recently shown that it is possible to use fMRI to investigate some features of orientation columns in humans. Rather than using typical fMRI data analysis in which each voxel is treated as a separate entity as far as statistical analysis is concerned, Kamitani and Tong (2005) used multivariate techniques and statistical pattern recognition algorithms to learn and later classify multivariate data points based on statistical regularities in the data set. Essentially, pattern-recognition algorithms operate by dividing a high-dimensional space into regions corresponding to different classes of data. This and other multivariate approaches are powerful because they can potentially discriminate between different classes of multivariate data even when the data, as projected along any one dimension, are indistinguishable. Kamikani and Tong (2005) used these techniques to show that distinct patterns of fMRI activity are produced by looking at differently oriented line gratings even though orientation columns are significantly smaller than the size of a voxel. By showing that there are small but stable biases[136] in the hemodynamic response of individual voxels to specific orientation patterns, they were able to use the information from these weak signals in many voxels (each of which, individually would not be sufficient to discriminate between different stimulus orientations) to identify distinct patterns of activity that correspond to different orientations. These patterns could also be identified in new data sets and reliably used to predict what stimulus orientation the subject was viewing. Thus, the granularity of the representation in this case is not to be understood as individual pixels, but as patterns of many (up to 100) pixels. The fact that an individual object of interest, such as an orientation column, may have a

[136] They suggest that such biases may arise from variability in the distribution of columns or of their vascular supply (2005, 5).

granularity below the resolution of an instrument, does not automatically mean that there can be no granularity match. If a question is directed at distributed patterns of activity among *many* small, sub-voxel size objects, then the granularity of the representation may still be sufficient to provide the relevant discriminations.

Reliability, then, refers not to the resolution of the instrument but to the relation between the granularity of the objects[137] in the world that are required to discriminate between the possibilities under consideration and the granularity of the instrument. Both the resolution of the representation and its granularity will be set, independently, by error bounds. Reliability comes in degrees and the degree of reliability is determined by the maintenance of structural features of the object in the data and by the extent to which the granularity of the world (the level of granularity at which the question of interest is directed) is similar to the resolution of the instrument or data.[138] I will refer to this similarity as a granularity match. For there to be a granularity match[139], however, does *not* require matching in the sense of an isomorphism or 1:1 correspondence between the structure of the world and that of the representation. Rather, what

[137] Or, more precisely, of those properties of the object or event that get represented.

[138] As was the case with reliability, I take the resolution of the data to be determined by the resolution of the instrument that produces them. If the question of interest concerns not the immediate input to the system – e.g. the source of 511keV photons – but some event such as neuronal activity that occurs further upstream, then the degree of resolution that characterizes the relation between this event and photon production will contribute to the relevant resolution of the final PET data for the purpose of addressing that question.

[139] Poeppel and Embick (forthcoming) discuss what they term "the Granularity Mismatch Problem" between neuroscientific and linguistic investigation of language. The idea is that basic linguistic concepts are usually more fine-grained than basic concepts in neurobiology and cognitive psychology creating difficulties for developing and testing hypotheses bridging the two domains. They also suggest a solution to the problem: we need to describe linguistic processes in computational terms at an appropriate level of abstraction, which they claim is that of neuronal populations (forthcoming, 4-5). In this way, we can connect linguistics with neuroscience. When I refer to a granularity match, however, I am referring not to the match between two conceptual domains, but between the object and the representation. Since the sort of representation that we can get using PET is quite heavily constrained (we can change many things about the visual display, but not about the effective resolution), we cannot simply re-describe or re-represent the object in order to get a granularity match. Neither can we change the granularity of a particular object: if a group of neurons that is involved in a particular task (e.g. an orientation column) is a particular size, it just *is* that size. We can change the question to be about objects with a different granularity, of course, but this obviously doesn't get us an answer to the question we started out with. Thus, what I mean by a granularity match (or mismatch) is considerably more difficult to alter than the sort of mismatch Poeppel and Embick discuss.

determines the presence and goodness of a match is the discriminatory ability that a particular representation provides relative to the discriminatory requirements of the question to be answered. The discriminatory ability of a representation is determined by the error bounds that apply to the represented quantities: as long as there is a sufficient agreement between the targeted granularity of the world and the (non-overlapping) error bounds of the representation[140], then the data allows the relevant discriminations to be made and are reliable for that purpose. If, on the other hand, the granularity of the representation is lower than the granularity of the world that is addressed by the question, then it cannot reliably discriminate between the relevant possibilities.

This account of reliability in terms of a granularity match improves on accounts of representation that require an isomorphism between the structure of the target and that of the representation (e.g. Cummins 1996; Giere 1998; Ziman 1978; Turnbull 1989) since it clarifies how a PET image or other representation can be used to make very reliable discriminations even in the absence of a 1:1 mapping. While accounts based on isomorphism generally deny that X can represent Y in the absence of a structural isomorphism between the two[141], all that is required on the granularity account is that the structures of the target that are relevant to a particular question be mapped somewhere within the error bounds in the representation. To see how this works, consider Figure 4.1. Panel A show what a traditional mapping account would require of a reliable mapping. The yellow oval on the left refers to some region of the brain activity. The smaller red circles are areas of activity, e.g orientation columns. The blue oval on the right refers to the granularity of the PET image (set by the

[140] Recall that it is not always possible to determine the error bounds. If the error bounds are not calculable, neither is the granularity of the representation.
[141] Cummins, for instance, defends a picture theory of representational content that requires strict isomorphism between object and representation.

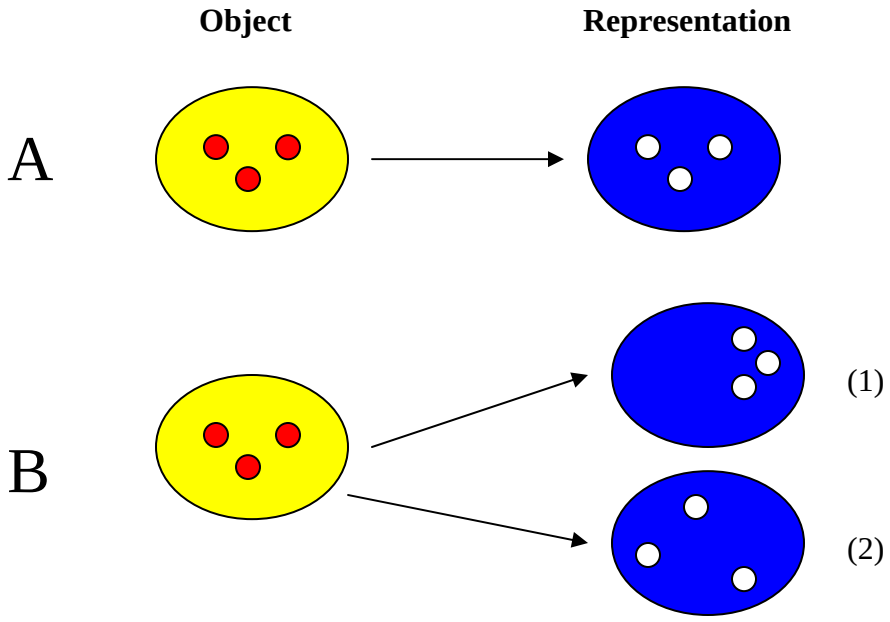


Figure 4.1 Granularity vs. a traditional mapping account.

error bounds) and the white circles represent areas of activation. A 1:1 mapping would require that the areas of activation stand in the same relationship to each other and the outer bounds in both the object and the representation. Panel B shows what is required given my account of reliability as a granularity match. There must be a correspondence between the granularity of the object and the granularity of the representation (the yellow and blue ovals), but differences that occur in the representation below the level of granularity need not stand in any single kind of mapping relationship. Thus, both (1) and (2) constitute a reliable representation of the real world object.

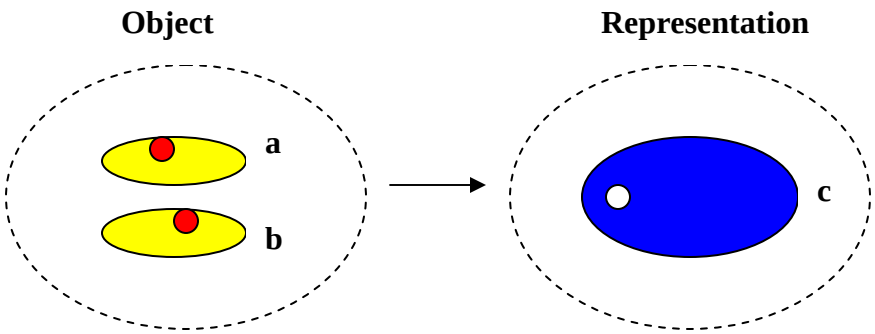


Figure 4.2 Absence of granularity match.

If, however, there failed to be a granularity match between the object and the representation, then no reliable information could be obtained. This is shown in Figure 4.2 in which the granularity of the objects in reference to which some discrimination is to be made is smaller than the granularity of the representation. If we wanted to know whether activity occurring with some particular task were occurring in area **a** or area **b**, but the granularity of our imaging technology were such that it could only tell us that it occurs within **c**, then it does not allow us to reliably discriminate between the relevant possibilities. This might occur, for instance, if we wanted to answer questions about the activity of isolated orientation columns using PET or fMRI. While identifying distributed patterns of activity increases the effective granularity of the representation (give the use of appropriate statistical techniques), the granularity of an individual orientation column is below the resolution of the instrument and, in this case, the granularity would not have greater dimensions than the resolution.

4.4. How can reliability be assessed?

I will now turn to the question of how reliability of a process (whether human perception or an instrument) can be assessed. The reliability of an imaging technology will be separated into the reliability of the instrument itself (for detecting what it actually detects, e.g. the location of annihilation events in the case of PET) and the reliability of the process of detecting the phenomenon that the data will be claimed to represent (e.g. the location of cancerous lesions or brain activity). To do this, it will be helpful to disentangle the notions of reliability and validity. While validity is usually subsumed by reliability in the epistemological literature, the difference is often important.[142] When claims about the validity of results obtained with some imaging technology are made or disputed, it is usually reliability in the second, broader, sense that is intended.

Bogen (2001, 2002) has argued convincingly that functional brain images[143] are better in the sense of being more reliable, than any current epistemological theory would allow them to be. Neither any traditional empiricist account, nor Mayo's error statistics, nor Woodward's counterfactual approach are able to explain why functional images seem to be as good evidence as they sometimes are. One difficulty for the anthropocentric empiricist is that no human observer perceives either the signals (i.e. positron-emitting isotopes, or, I will claim, photons, in the case of PET) that are detected by the instrument, the physiological phenomena that are presumed to be causally related to the distribution of the signal within the brain (increased blood

[142] The difference may not be as important in the case of human perception or in thought experiments where validity is assumed to – and usually does - accompany reliability. However, the difference makes a difference in the case of imaging technologies since they may more often produce very reliable data that systematically misrepresent (or even fail to represent) the phenomenon of interest.

Interestingly, in some work on validity within philosophy of science, the opposite holds and validity either assumes or subsumes reliability

[143] He discusses both PET and fMRI, but I will continue to restrict my discussion to PET.

flow), or the phenomena of interest itself (cognitively significant brain activity) (Bogen 2002, S60). Another problem for an anthropocentric empiricist account is that the intensive statistical and mathematical processing required to produce this sort of data seems to irremediably blur the distinction between producing and interpreting data (Bogen 2001, 174). On these points, Bogen is entirely correct and will not say any more about them than I have in earlier chapters. I am also largely in agreement with him as far as his characterization of the inability of both Woodward's and Mayo's accounts to allow for the epistemic quality of functional brain images, though I think that we can do a substantially better job of assessing the error characteristics of PET than he allows. More importantly, my account here has aimed to provide a much more substantial account of what it means for this sort of evidence to be reliable. In addition, in examining the ways in which we can assess reliability I claim that we can and should assess the reliability of PET not only as an instrument that detects brain activity, blood flow, or cancerous lesions, but also as an instrument that detects the spatial and temporal source of 511keV photons within an object. In essence, what we need to do is to isolate the instrument itself from the upstream and downstream processes that are, variously, added to the instrument when it is used to try to detect different phenomena of interest. If we do this, we can provide a better account of the reliability of the instrument itself than is apparent on Bogen's characterization. This is made possible by the fact that there are far more empirical tools available to test different aspects of PET, including the algorithms, when we look at medical and specifically oncological applications. Once we can improve our account of the reliability of the instrument, we can then focus on trying to assess the reliability of the upstream and downstream processes that flank the instrument in the experimental set-up used in investigating particular types of phenomena.

The first thing to do, therefore, is to set out what I take to be the bounds of the instrument. In keeping with the last chapter where I discussed human perception and PET partially in terms of input and output, I will again take the input of PET to be high energy (511 keV) photons and the output to be either the reconstructed, attenuation-corrected image or the numerical vales for each voxel that correspond to the image. For the present purpose, the two are functionally equivalent since going from the numbers to the image is purely conventional and is as close to perfectly reliable as any process can get.[144] I differ from Bogen in taking the input to the instrument to be photons rather than radiation (positrons). Because he is concerned with the process of detecting brain activity as a whole, however, this difference has no real significance for his account. With respect to the anthropocentric empiricist position, for example, it makes no difference whether it is positrons or 511keV photons that get detected, since human observers are incapable of detecting either. Similarly, we are no better or worse able to identify counterfactual dependencies or statistical error for one relative to the other. Since it is close to guaranteed that an emitted positron will collide with an electron (and so produce a pair of 511keV photons) after traveling less than 1mm, this distinction actually makes virtually no practical difference to my account either. However, because it is the photons and not the positrons that actually strike the PET detectors and are counted, I do want to insist on taking photons to be the input to the instrument. Annihilation events and positron emission are the immediately proximal upstream events and are characterized by highly reliable physical interactions, but they are part of the upstream sequence, not the input.

[144] Recall that a simple thresholding method is used to assign color to each voxel in an image. Any voxel that has an intensity value with a specific range will be shown as a particular color in the image. The only sorts of errors that could occur are programming errors that result in faulty assignments. In Chapter 5 the distinction between the image and the numbers will become important, however.

With these boundaries in place, it is obvious that applications of PET for both brain imaging and oncology can share an assessment of reliability with respect to the instrument and the immediate upstream events of positron emission and annihilation events. Any method that can be used to establish the reliability of different arrays of detectors, different scintillating crystals, and, more importantly, different algorithms for noise elimination, image reconstruction, or attenuation correction, can tell us about the reliability of the instrument portion of the experimental set-up for either sort of application. This is very important since there are many more options available for checking the reliability of PET in oncology than there are in neuroimaging. For instance, the existence of a tumor in the identified location can be checked by biopsy and histology.

At this point I need to review a terminological matter than was mentioned earlier only in passing. I have been using the term reliability in accordance with common usage in epistemology and much of the philosophy of science to refer not just to repeatability of data gathered using some method, but to the idea that the data actually reflects the phenomenon of interest, not interfering factors of various kinds (i.e. it is not artifactual). The latter notion, however, might more properly be referred to as *validity*, and specifically as internal validity. Allan Franklin (1989), for instance, has identified a set of epistemological strategies for establishing the validity – primarily in the sense of internal validity – of an experimental result or observation. Franklin's strategies also assume that results are reliable in the sense of repeatable, but since this is a far easier thing to establish, he likely never felt the need to point out the fact that you need reliability too. While it may introduce some confusion in the minds of readers more familiar with the proper use of the term "validity", especially as I am about to discuss some

of Franklin's strategies, I am going to continue to use the term "reliability" for what we seek from instruments or experimental processes in order to establish their epistemic credentials.

It will be helpful to consider how the reliability of PET can be assessed in terms of the strategies that Franklin (1989) has suggested are used to assess whether the results from an experiment are reliable (in my sense). They can be used to establish, in other words, that the instrument is working properly and that the data accurately[145] reflect features of the phenomenon under investigation. They are not intended to be either individually or jointly sufficient to guarantee the reliability of observational data, neither does he take the set to exhaust the possible strategies. Nevertheless, they provide a broad sample of the sorts of strategies that can be used to help establish reliability. Franklin's examples come primarily from material experiments in physics, but Rudge (1996, 1999) has shown that the same sorts of strategies are used in evolutionary biology and Parker (2003, unpublished manuscript) has demonstrated that they can also be used to establish the reliability of computer simulation experiments. It is not my intention to demonstrate that each of Franklin's strategies can be used in the case of PET, but rather to use his strategies as a framework for highlighting particularly important methods for determining the reliability of PET. Accordingly, I will not even mention some strategies and will deal with others very briefly while discussing some at greater length.

4.4.1. Strategies for assessing the reliability of PET

Replication of results using a different apparatus.

If the same phenomenon can be observed using a different apparatus, especially one that is based on very different physical characteristics (e.g. Hacking's example of using different types of microscopes to establish the existence of dense bodies in cells), then we ought to have increased confidence that when we observe the phenomenon using the instrument whose

[145] With a degree of accuracy sufficient to make the discriminations required.

reliability is in question, it is a real effect. The difficulty with PET, as Bogen (2001) points out, is that the only other tools we have right now for observing the same sorts of cognitive activity as PET rely on many of the same statistical methods as well as physiological assumptions. Thus, these methods are not different in a very significant way. However, we can observe cancerous lesions that are identified by PET in other ways: by biopsy and histology, as well as, in the case where histology (generally believed to be the gold standard for detection of cancer) has confirmed the presence of an operable lesion, by surgery. Both biopsy and surgery provide naked eye confirmation of the location of the tumor; surgery also confirms its size and shape. Histology provides an interesting contrast since it detects not gross morphologic features of a tumor, but changes at the cellular and subcellular level. This provides support not only for the reliability of the PET instrumentation (everything that lies between input and output of the instrument), but for the physiological changes that are associated with cancerous cells and that the radiopharmaceuticals used for oncological applications (FDG-glucose, as well as less commonly used compounds) are actually identifying individual cells with metabolic and other changes associated with cancer.

Indirect testing

This is a potentially very valuable technique for determining the reliability of PET since it is a strategy that can be used when a particular observation (e.g. of brain activity in the intact living human) can only be made with one kind of instrument. In practice, however, its usefulness is relatively limited. If we can observe a particular phenomenon, p_1, only with one instrument but we can use that instrument to observe another type of phenomenon, p_2, which can also be observed using some other instrument, then replicating our observation of p_2 in the second instrument serves not only to establish the reliability of our observation of p_2 using the first

instrument, but helps to establish the reliability of the first instrument for detecting p_1. Thus, our ability to use biopsy, histology, and surgery to help establish the reliability of PET for detecting cancer also helps to establish its reliability for detecting brain activity. Its potential value, however, is limited to establishing the instrument itself and the immediate upstream steps of positron emission and annihilation events. Because the physiological events or features that lie further upstream are not at all related in the case of oncology (e.g. increased glucose metabolism of tumor cells) and functional brain imaging (increased blood flow in response to an increase in neuronal activity), this strategy cannot provide any support for the reliability of these processes and of the considerable role they play in the overall reliability of PET for functional brain imaging.

Intervention and prediction

This is another strategy that can be used in the case of phenomena that can be observed using only one technique and is perhaps the most valuable technique that is available in the case of PET. It involves controlled manipulation of the observed objects and determining whether or not the instrument gets the predicted or correct results. Ethical as well as practical considerations preclude the sorts of experimental interventions on living humans that would be needed to test the use of PET for either oncology or brain imaging. However, so-called "phantoms" that have known characteristics can be specially constructed for both sorts of applications and the reliability of different sorts of statistical methods (within the PET instrument itself) compared with respect to how closely the results of using them match the known features of the phantom. The use of phantoms in oncology was discussed in Chapter 2 in the context of the debate over whether attenuation correction increased or decreased reliability. To review briefly, a phantom is a physical model of some part of the human body or of a simpler geometric form the

composition of which is known. The relevant features will normally include the position, shape, size, radiation concentration, and attenuation coefficient of each compartment. The phantom is scanned under different conditions and the resulting data compared to its known characteristics. For instance, Hsu (2002) used an anthropocentric thoracic phantom into which was placed (in the left breast) a "lesion" with a radiation level five times background and a volume of 2 cm^3. The phantom was scanned and the initial image reconstruction performed using four different algorithms. By comparing the data obtained using each algorithm and comparing it to the actual features of the phantom, Hsu was able to determine not only which algorithm was the most reliable, but also to get information about how different algorithms performed over time (of two iterative algorithms that provided equivalently good data in the end, one required 4 or five iterations to reach what was deemed to be a reasonable level of convergence, while the other took 20 iterations and was four times slower to achieve a similar convergence).[146] Phantoms, therefore, are an invaluable resource for testing the reliability of various statistical techniques.

Phantoms are also used to validate or optimize PET methods for brain imaging. In this case, however the phantoms used are not physical models as described above for oncology, but digital models. They consist of multiple data sets that are defined in terms of voxel size and intensity. The phantom can be used to simulate different patterns of brain activity by creating data sets that vary in the location and size of areas of increased voxel intensity (e.g. Schoenahl et al. 2003; Lukic, Wernick, and Strother 2002; Collins et al. 1998; Hoffman et al. 1990) These digital phantoms do not actually get scanned by the instrument as do the radioactive phantoms used to evaluate algorithms for use in oncology; they can only be used as input data to the computer to test how different algorithms will perform. However, they serve a very similar role in allowing

[146] Computational time required is a very significant consideration for clinical practice, as was suggested earlier in the discussion of logical vs. finite probabilistic reliability.

the performance of different algorithms to be evaluated in terms of how well the final data it generates correspond to known data sets. Since we have no other (sufficiently different) way of observing the sort of brain activity that we use PET to detect, using simulated data is the only way in which to empirically assess the performance of algorithms specifically for data of the kind to be expected in brain imaging (i.e. activations that are often smaller, more spatially distributed, and of much shorter temporal duration than phenomena observed in oncology).

Properties and theory of the phenomena

It may be the case that observed patterns in the data are too consistent and natural looking to make it plausible to interpret them as artifacts. This strategy does not seem to be particularly useful in the case of PET since there are so many sources of error that it is unlikely that no other plausible account could be given for any particular pattern observed. On the other hand, this seems to be a very common strategy for assessing the reliability of human perception.

A well-corroborated theory of the instrument

This is a strategy that can be used much more successfully imaging technologies of various kinds than for human perception. As the last chapter showed, our knowledge of the functioning of the visual system is still far too incomplete for us to claim to have a theory of it that can provide support for claims that it is reliable. On the other hand, we know exactly what sorts of physical and statistical processes contribute to PET and often, if not always, have a good idea of the sorts of errors that a particular algorithm will make. To take just a single example, identification of active voxels in neuroimaging data usually involves performing voxel-by-voxel statistical tests and setting some threshold according to which voxels will be classified as active (if above the threshold) or inactive (if below). The selection of thresholds that are both objective and effective in terms of limiting the false positive rate has been an enduring problem.

Theoretically motivated thresholds (e.g. always setting the significance level at the conventional >0.05) result in a very high rate of false positives since so many tests (28672 for each voxel in a 64 X 64 X 7 image) are performed. Standard methods for multiple hypothesis testing (such as the Bonferroni correction)[147] are often not sensitive enough for neuroimaging. It can be shown statistically that the Bonferroni correction tightly controls Type I error and, when applied to the entire data set, has a tendency to eliminate both true and false positives (Genovese, Lazar, and Nichols 2002). More complicated methods can be used, but these usually require either increased data or increased computational time and so are not always feasible. Another alternative it to reduce the number of comparisons that are performed simultaneously, for instance, identifying regions of interest and applying the correction to each set of voxels separately. However, in order to be objective, regions of interest must be created prior to data analysis and must be left unchanged, a condition which often proves to be too rigid. One alternative strategy that has recently been proposed is the use of procedures that control the false discovery rate (FDR). According to Genovese, Lazar, and Nichols, "the FDR is the proportion of false positives (incorrect rejections of the null hypothesis) among those tests for which the null hypothesis is rejected. We believe that this quantity gets at the essence of what one actually wants to control, in contrast to the Bonferroni correction, for instance, which controls the rate of false positives among all tests whether or not the null is actually rejected." (2002, 871).

[147] The Bonferroni correction adjusts the level of statistical significance that is required to reject the null hypothesis according to the number of tests that are performed. Essentially, it decreases the significance level (*p*-value) that is required for an individual test so that type I errors are reduced and the study-wide significance level remains at <0.05. The general formula for the adjusted significance level is $1-(1-\alpha)^{1/n}$, often approximated as α/n. To see why this helps, consider that if the null hypothesis is true and the significance level set at $p>0.05$, a significant difference will probably be observed by chance once in every 20 trials. So if you were trying to assess the activity status of 20 voxels (far, far fewer than any actual PET data set), and the null hypothesis were, in fact, to hold for all of them, the chance that at least one of them would (incorrectly) be judged to be active is not 0.05 but 0.64. Applying the Bonferroni correction to this case, the significance level for each test would be set at 0.00256.

A well-corroborated statistical theory can help with the statistical aspects of the instrument (as well as any statistical steps located upstream or downstream of the instrument), but other causal aspects of the experimental set up require different corroboration. The connection between increased glucose metabolism and tumor cells is something that has been well-established by physiologists and cell biologists. The connection between blood flow and brain activity is also generally accepted to hold. However, it is also widely acknowledged that neuronal activity occurs on a much smaller time scale that does a change in blood flow. Thus, knowledge of the experimental set-up (of upstream events in this case) tells us that any brain imaging technique that relies on blood flow to indicate cognitive activity will have too low a resolution to make discriminations for many questions of interest. (Recall, however, that low resolution does not imply low reliability).

4.4.2. Success of strategies for assessing reliability

While I focused in the above discussion on PET, I did make brief mention of human perception at a couple of points. This hopefully served as a reminder that while I hope to give an improved account of the reliability of PET based on a separation of particular applications into the instrument itself, upstream elements, and downstream elements, I also need my account of reliability to apply to human perception. While we do not generally have available any sort of error statistics for human perception under normal conditions, this need not undermine the account since Mayo claims that "in practice often informal and qualitative arguments may be all that is needed to approximate the severity argument. Indeed, perhaps the strongest severity arguments are of a qualitative variety." (2000, S202) She cites as a good instance of this, Hacking's argument for taking dense bodies to be a real rather than artifactual. This same argument was cited by Franklin as an instance of the replication of results using a different apparatus. If no direct appeal to any formal statistical model is required to run a severity

argument, then it seems plausible to claim that the other epistemological strategies that Franklin identifies may also count as qualitative severity arguments. In this way, we may argue for the reliability of human perception based primarily on qualitative arguments about the plausibility of the data obtained with it, our ability to intervene with the objects of perception and get the predicted results, our (admittedly non-technical) understanding of the instrument in terms of the conditions under which it works and those under which it doesn't, and our ability to confirm at least some observations (e.g. the shape, size, and texture of objects) by using other sensory modalities (especially touch).

Assessment of the reliability of PET will rely more heavily on actual statistical claims. I am more optimistic than Bogen (2001, 2002) about the ability of an error statistical approach (suitably fitted with a finite frequency interpretation of probability). In part, this is because I have ignored the downstream components of PET for brain imaging. While mapping images onto the standard Talairach-Tournoux atlas is still common[148] and is responsible, as Bogen claims, for introducing a significant amount error, there are a large number of alternative warping algorithms that can be used to perform this task and can be chosen with an aim to maximizing reliability for the particular biological question at hand. This area is very complex, however, and an adequate discussion would take a lot of space while contributing very little to my story. It will suffice to point out that both the linear transformations available with the Talairach-Tournoux atlas and the more complex warps now available (if not often used) can fit into my account of reliability. They may often be the resolution-limiting step and so reduce the number and type of question that the method can be used to address, but this does not necessarily

[148] Though the Montreal Neurological Institute (Evans et al., 1994) created a composite MRI dataset from 305 young normal subjects to deal with some of the concerns that had long been expressed with the actual template used by Talairach (post-mortem sections of a 60-year-old woman). The resulting average brain has some blurring of individual structures where spatial variability in the population is high, but the template is used with increasing frequency as part of the common Statistical Parametric Mapping (SPM) template.

indicate that they reduce the reliability of the overall process. This is not to say that they may not sometimes also decrease the reliability, only that they need not do so.

4.5. Conclusion

I have argued for an account of reliability that shares features with both Goldman's reliable process account and Mayo's error statistical approach. It captures Goldman's fundamental idea that a reliable process ought to be truth-conducive while adopting a more substantial account of what it means for a process to tend to produce more true than false beliefs. Importantly, the account is not a strictly causal account, thus, it can be applied to both the straightforward physical or causal processes that contribute to both human perception and imaging technologies and to the mathematical and statistical techniques that are used in various imaging technologies.[149]

My account also suggests solutions for some of the key objections that have been raised against reliabilist accounts in general. To recap, the objections and their solution were as follows:

- A reliabilist account must explain how a process is to be judged to be reliable. This was achieved by supplementing Goldman's basic intuition that a reliable process should produce a high ratio of true beliefs to false beliefs with a more precise account of reliability drawn from Mayo's error statistics.
- It must make sense of the idea that a process type can tend to produce data with good long-run error characteristics. An error statistical approach provided part of

[149] While neural aspects of visual perception are often referred to as computational or algorithmic, our current lack of knowledge about these processes means that we cannot hope to assess their reliability through examination of the computations or algorithms themselves. On the other hand, this is exactly what we can and must do in the case of imaging technologies.

the solution here but the sense of probability used had to be specified. Goldman wanted to understand probabilities as propensities and Mayo was vague on the issue, claiming only to want a frequency interpretation. It was claimed that the goals of objectivity and having the best chance of accepting the correct hypothesis in the (relatively) short run given a finite amount of data requires specifically a finite relative frequency interpretation.

- It must solve or avoid the generality problem. This problem was easier to deal with in the case of imaging technologies since we know exactly what sorts of physical, computational, and algorithmic processes contribute to them and since their input is heavily constrained. In the case of human perception, it was suggested that a similar account in terms of the computational and algorithmic processes involved might help overcome this problem.

In addition, my account proposed a distinction between reliability and resolution. Reliability, like resolution, can be had to greater or lesser degrees and discrimination between particular statements about the phenomena may require a method that has a certain minimum level of resolution and a minimum level of reliability. However, a method that has a lower degree of resolution does not necessarily have a lower degree of reliability.

Finally, this account, together with a distinction between the instrument itself and upstream and downstream components of the experimental set-up for specific applications of PET was able to be used to help improve our understanding of when and why PET data is reliable.

5. Why pictures?

5.1. Introduction

Imaging technologies seem, by their very name, to refer to the production of images yet it is a striking feature of many imaging technologies that their output need not be images. When we examine the means by which PET images are produced, for instance, we see that the fact that they even *are* images is accidental. While a photograph may be measured and subjected to quantitative analysis subsequent to its production, PET images require that extensive mathematical transformation occur to produce the data that can *then* be represented in the form of an image. In the case of PET, the result of signal detection, data correction and reconstruction is a numerical value assigned to each voxel. The final conversion of this data into the form of a vaguely naturalistic image is simply a matter of assigning a color (or grey level) to particular ranges of numerical values and then displaying the data in a 2-D or 3-D array. It could just as easily be represented in other ways. For instance, the change in the average voxel intensity within some defined region or regions of interest over time could be displayed in graphical format. Some neuroscientists and cognitive scientists, in fact, prefer to represent their data in this way.[150] Yet, every scientific paper that reports data from functional imaging studies contains at least some photograph-like images.[151] Given that there is a choice between data display formats, why are images the dominant form? A full answer to this question clearly involves historical, sociological, and rhetorical perspectives in addition to the epistemic one. In this chapter, however, I will only be able to briefly identify a few of the historical and other

[150] Julie Fiez, personal communication.
[151] At least, I have not been able to find any that do not.

features that may contribute to the dominance of images. My primary concern will be with the question whether there is an epistemic advantage to using images. I will examine two advantages that might be claimed for images - cognitive accessibility and facilitating the identification of causal relationships (in the case of kinetic images) – and argue that only the first actually holds.

5.2. What can we see in the data?

Some types of data often seem to us to essentially wear their reliability on their sleeve. We tend, for instance, to take photographs, video recordings, and photograph-like images such as X-rays as reliable forms of evidence for certain sorts of visually accessible features of the world.[152] This is, in part, a consequence of the fact that the processes involved in producing these forms of evidence usually *are* reliable. However, it is also partially explained by our familiarity with these sorts of images and the sorts of things they often represent. We are all highly trained in reading these *types* of images, even if we are not always expert in identifying or interpreting specific kinds of content. The layperson looking at a photograph of a face, for instance, will recognize it as a face though that same person looking at a photograph of a tissue sample stained to reveal macrophages may have no idea what they are looking at.[153]

This familiarity with certain visual formats not only allows us to identify their content, but, importantly, also often allows us to judge the *reliability* of the image. We know both what a

[152] However, Daston and Galison (1992) point out that what counts as objectivity has changed over time and is reflected in the practices of scientific image-making. It is not always clear which objects, if reliable images are produced of them, serve as reliable information about some phenomenon or feature of the world. For instance, in examining the number of immune cells of a particular type that are present in different layers of the skin in normal as opposed to scar tissue, should I always photograph and count random fields of view or ought I to instead require that fields be randomly selected but meet some additional criteria – perhaps that they not contain any tears or that the tear not cover more than a certain percentage of the area of the field. Some such criteria undoubtedly enhance the reliability of the data, but it is not obvious just what sort of constraints are there on the criteria that count as legitimate.

[153] This need for some knowledge or interpretive framework to see something *as*, for instance, a face, rather than just seeing a mixture of different colored areas was noted by Hanson (1965). Kuhn (1970) also discusses seeing versus seeing as in the context of the theory-ladeness of observation.

(real) face looks like and what a reliable photograph of a face looks like. Some sorts of variation we know to be permissible – we do not think that a black and white photograph is generally unreliable, for instance.[154] However, if we see a very blurry photograph, we will be more inclined to question its reliability because we can tell that something has gone wrong in the production of the image. Our ability to read the content of certain types of visual images is sometimes but not always connected to our ability to read their reliability. This is very evident in the photographic representations of visual illusions such as the Ames room in which people of the same height appear to be dramatically taller or shorter than each other depending on their position in an oddly shaped room that appears perfectly rectangular from the limited perspective of the viewer.[155] In this case, the unrecognized unreliability is created by the absence of some depth cues, not from unfamiliarity with what people of different height look like in a "normal" photograph. In other cases, we may be relatively unfamiliar with the content of an image and still be able to recognize it as less than maximally reliable. In the case of the tissue sample, for instance, some people would likely still be able to identify blurriness as a problem, though they would probably not pick up on other problematic features related to content and to parts of the experimental set-up upstream of the production of the photograph.[156] Thus, while it is obviously not always correct to do so, is easy for us to interpret reliable-appearing photographs or photograph-like images[157] as being, in fact, reliable.[158]

---

[154] It may be unreliable for certain purposes, but even then we can usually identify the purposes for which it is unreliable (e.g. for discriminating between green and blue eyes).
[155] See Crick (1994, 45-6).
[156] For instance, they would be unlikely to identify a photograph where all the cells in the tissue were uniformly stained as indicating a problem with the staining technique.
[157] Hereafter I will use the term "photograph" to refer to actual photographs as well as other images such as X-rays that are produced by processes that bear a physical similarity to optical photography.
[158] An additional difficulty in identifying photographs as reliable or not is digital manipulation of images after their initial production. This is a very real concern today given the ease with which images can be altered and has been addressed in several recent pieces in *Science* and *Nature* (Ottino 2003; Pearson 2005; Greene 2005). Part of the difficulty is in establishing what degree or type of manipulation is legitimate - i.e. does not compromise the

This connection between our expertise in reading photographs and our inclination to interpret them as being reliable is undoubtedly connected to the persuasive power of images, a topic that will be briefly returned to in section 5.3. It also highlights the importance of distinguishing between the *production* of data (the relationship between the object and the representation) and the *use* of data (the relationship between the representation and the human user – or observer – of that representation). My focus up to this point has been primarily, though by no means exclusively, on the former. However, the fact that the same data, obtained by a specific process for a particular sample, can be displayed in a variety of different ways now requires more attention. It is an important feature of the use of data that different representations of the same data may be interpreted very differently by the user.

Recall that in order for an image or any type of data to be reliable, it must satisfy the two criteria described in the previous chapter. First, there must be a granularity match between the instrument and the description of the world at which a particular question is directed. Second, the structure of the object must be preserved in the data within finite error bounds. Whether these criteria are met in any given case is in part dependent on the question to which an answer is sought, but the range of possible questions that can be reliably answered using a given instrument is itself constrained by the nature of the processes involved in the instrument.[159] This account of reliability does not distinguish between numerical data or images: both can be described in terms of error bounds and granularity.

---

truthfulness or reliability of the data and may, in fact, aid the viewer in making relevant discriminations - and what constitutes fraud or misrepresentation of what was originally perfectly reliable data. These sorts of conditions on selection and manipulation of data, however, are not specific to images but apply to all sorts of data production methods.

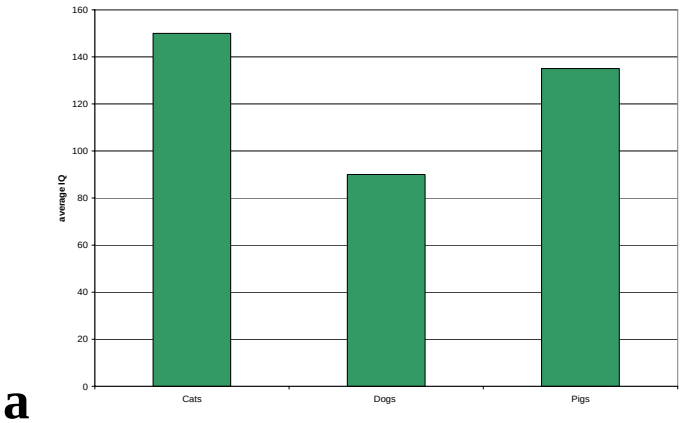
[159] More accurately, this constraint is enforced not only be the instrument (which has defined start and end points, as indicated in Chapter 4 for the case of PET), but by the experimental set-up including elements upstream of the instrument.

An essential part of reliability, though, is allowing certain objects, properties, or features to be discriminated. The ability of the user of the data to make certain types of discriminations is affected both by the data production process (constraints imposed by features of the object-representation relationship) and by the data display format. Features of the data production process obviously limit the types of discriminations that can be made using a given instrument. For instance, if one object contains twice the radioactivity as a second, but my method of data production cannot discriminate between x and $2x$ over this range of radioactivity, the same value will be assigned to both objects and I will not be able to discriminate between the two (at least not with respect to their level of radioactivity). Situations such as this can arise from either the detector used by the instrument or from the mathematical or statistical processing. If a particular detector can (or is set to) collect only a specific range of wavelengths or can register only a maximum number of radioactive counts per second, then it can obviously not be used to make discriminations that would require data about other wavelengths or distinguishing between two different radiodensities both if which produced counts above the maximal rate. Alternatively, mathematical or statistical features such as partial volume effects (see Chapter 2) may be the limiting factor.

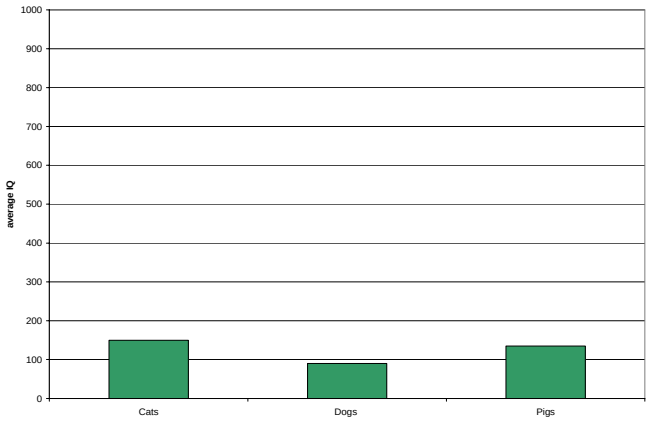
However, it is also the case that the way I choose to display the data can make it much easier or much harder for the set of possible discriminations to actually be made by the human viewer of the data display. Tufte (1983, 1997) provides a guide to the types of visual display that allow human user to make more discriminations and to make them more easily. But, as Tufte also notes, there are also formats that inhibit our ability to identify specific features of the data. Examples where the viewers' interpretations of the data are manipulated, intentionally or not, by the choice of graphical method are very familiar. For instance, choosing a larger or smaller scale

for the y-axis of a graph can make two quantities appear much more similar or dissimilar than they would had another scale been used, as is shown in Figure 5.1. The same holds true of photograph-like images. Figure 5.2, for example, shows 41 images that all represent the same PET data but assign different colors to different ranges of voxel intensities (the first and last image are shown with a linear grey scale). Notice in particular how the spot at the base of the image appears hot, cold, or even absent according to which color scale is used. Similarly, in Figure 5.3 the use of a pseudocolor palette in the images on the right allows us to more easily discriminate different voxel intensities and makes the difference between the top and bottom images appear to be much greater. Thus, the potential advantage of images over other data formats is highly dependent on specific features of the image display. For the remainder of this chapter, therefore, I will assume that images and other forms of data display are created in such a way as to maximize the ability of the user – *given a particular display format* – to make correct discriminations.[160] Notice that this does not prejudge the question of whether a particular display format has an epistemic advantage relative to other display formats; it merely recognizes that for each display format there are ways to increase or decrease its effectiveness. Just as the choice of color palette for an image may be optimized for making the discriminations of interest in a particular case, so too may there be optimal graphical or numerical representations of the same data. Presenting only a subset of the data in graphical format, for instance, can make it easier to identify features relevant to a making a given sort of discrimination.

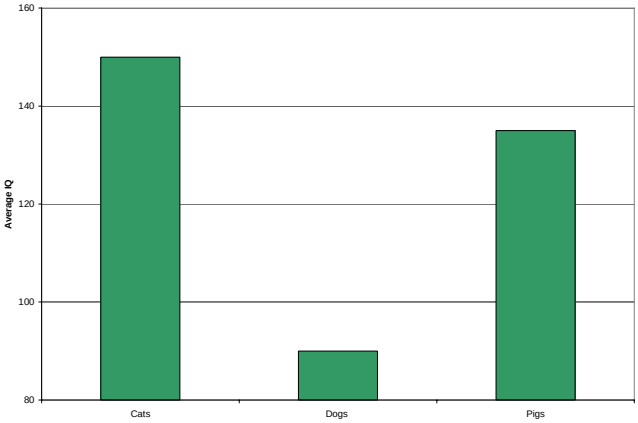
[160] "Correct" here refers to the discrimination(s) needed to answer the question of interest in any given case. It will sometimes be the case that display formats that make some features of the data more easily discriminable by the user also obscure or make impossible to discriminate other features. For instance, if what is needed to answer a particular question is the ability to discriminate relatively small differences within a specific, limited range of intensity values, a different pseudocolor may be assigned to small intensity intervals within this range and larger intervals outside of it. This will, in effect, visually eliminate some differences that occur outside of the intensity range of primary interest. This is, of course, purely a matter of the representation-user relationship, the differences are not eliminated from the numerical data (the object-representation relationship is unchanged by this sort of manipulation).



a



b



c

Figure 5.1 Visual effect of changes in graphical scale.

All three graphs show the same data for the average IQ of three different animals: cats 150, pigs 135, and dogs 90. The only difference between the three graphs is the choice of y-axis. In a, it starts at 0 and goes up to 160. While cats and pigs are clearly smarter than dogs, the dogs don't appear to be too badly off. In b it starts at 0 and goes up to 1000. All three animals seem to be almost equally mentally challenged. In c the y-axis begins at 80 and goes up to 160. Cats and pigs again seem to be much smarter than dogs, which now appear to have some real difficulties.

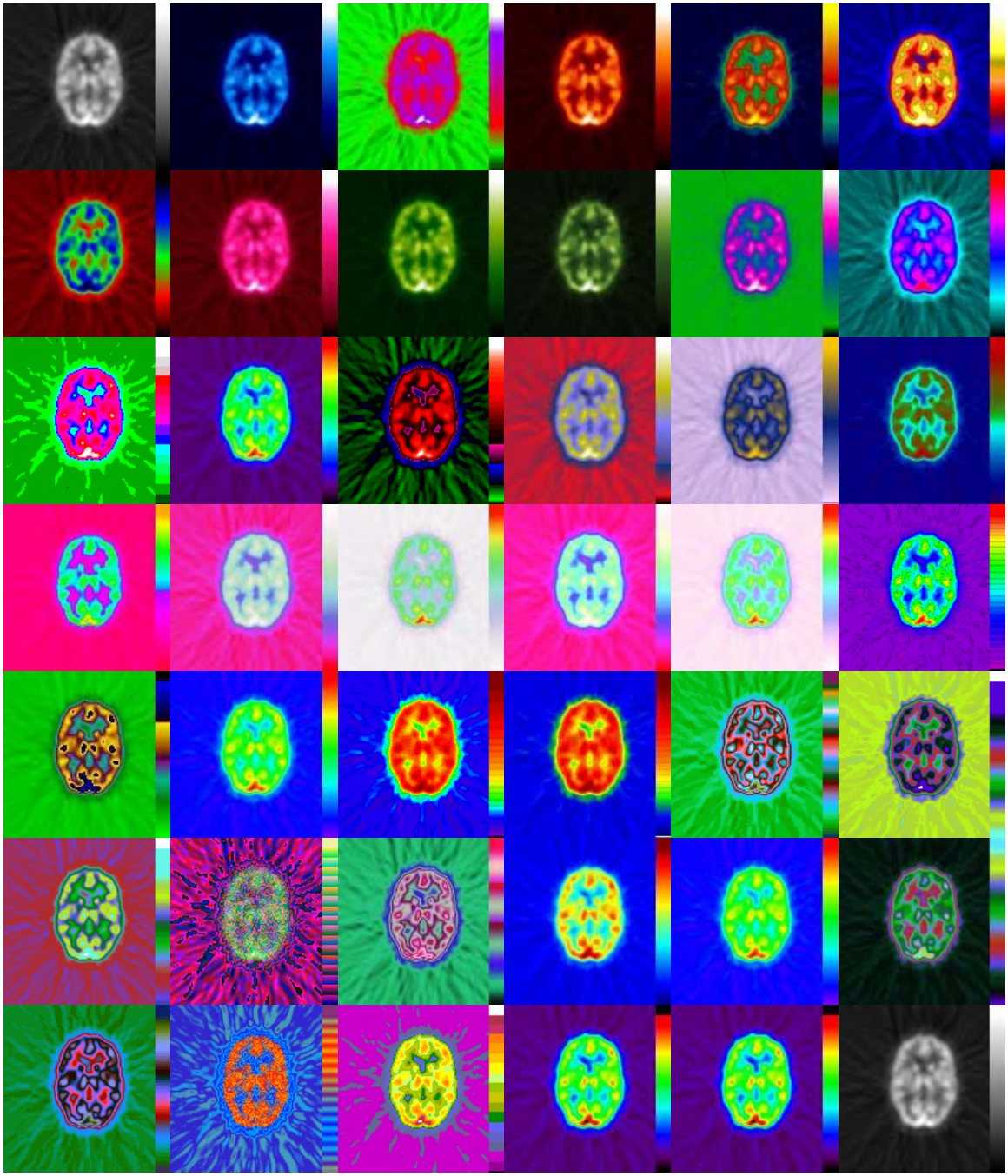
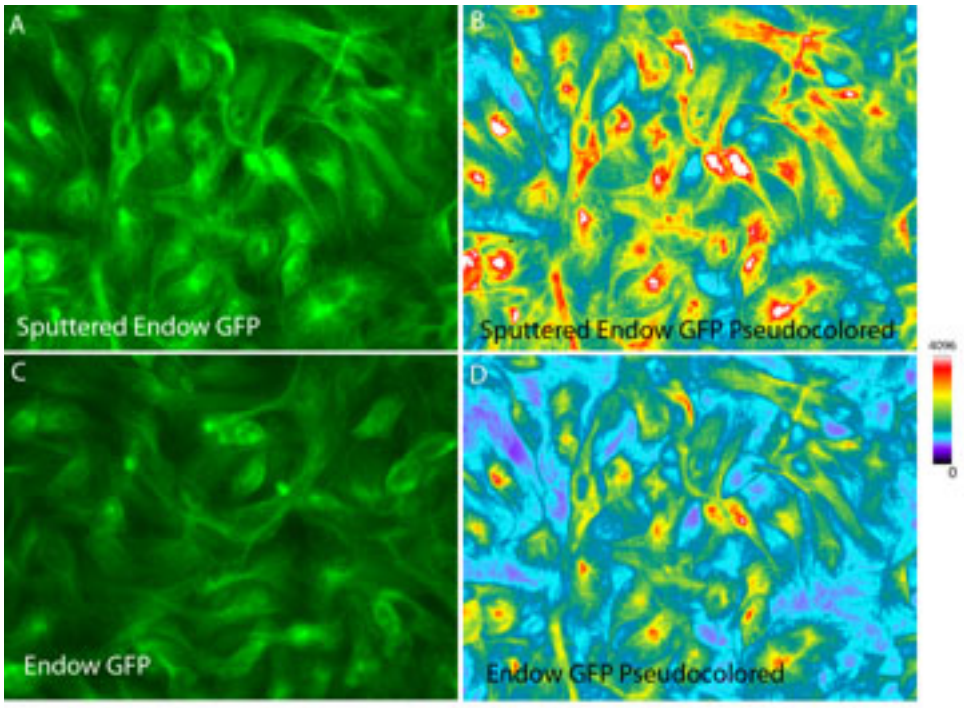


Figure 5.2 Identical PET data displayed using different choices of color scale

Original images taken from Brian Murphy (1996). This series of images appeared on the cover of the December 1996 issue of the Journal of Nuclear Medicine Technology and was intended to demonstrate the effect that choice of color scale has on how we "see" the data.



Intensity comparison between conventional (C, D) and Sputtered (A,B) Endow GFP Cubes. (60X 1.45NA, GFP Actin)

Figure 5.3 Effect of pseudocolors

Demonstration of the difference in visual effect of the use of pseudocolors rather than variation in brightness of a single color to represent intensity differences. (Images courtesy of Simon Watkins)

For the purpose of examining whether visual images have any kind of epistemic advantage, it will be important to keep in mind not only that any display format can be used more or less effectively, but that, in the case of the sorts of imaging technologies that are the focus of this dissertation, the use of different data display formats does not indicate a difference in the object-representation relationship, but only in that between the representation and the user. The data collected using the instrument is the same no matter what form of data display is chosen.[161] The

[161] This will not be true of all other types of instruments. An instrument that uses X-ray or photographic film as the detector, for instance, does not first represent the data in numerical form. The image format in such cases is not optional in the sense that it is with something like PET. The data can be converted to numerical format (e.g. by

numerical value associated with each pixel or voxel is not changed when we represent it in a different format or by a different color within a certain format. Thus, if images are to provide some kind of advantage, it will be in terms of their use by the viewer rather than in terms of their content.

Given the above, images are potentially able to play two important epistemic roles.[162]

- The first is *cognitive accessibility*: images make many features of the data set (overall patterns, relationships between parts of the images and between large and small scale structure) more easily accessible to the human cognitive system than do other types of display such as linear strings of numbers. This is particularly true for very large, complex data sets such as those produced by PET and confocal microscopy. It might well be possible for me to extract as much information from a string of numbers identifying the number of blades of grass in each of two halves of a 1 square inch patch of lawn as it would be to see an image of the area with different colors used for different numbers of blades, but as the number of data points increases, so does the efficiency of the visual over the numerical display. A graphical representation of the patch of lawn divided into quadrants rather than halves would, at least for most people, probably make it easier to identify the spatial relationship (directly vertical, directly horizontal, etc.) between the most grassy and least grassy quadrants. For the tens of thousands of voxels in the average PET image, there is no question of our being able to identify areas of high of similar activity by looking at strings of numbers, let alone being able to tell what region of the brain those areas

---

scanning or otherwise digitizing the image) and then represented in other formats, but in this case it is not strictly accurate to claim that the *same* data is displayed as an image or in other forms.

[162] Wimsatt (1991) similarly identifies visual representations as the simplest and most inferentially productive means of analyzing multidimensional data and processing information about motion.

correspond to. But this information is very readily picked up by even a quick scan of the PET image.

- The second potential role is in facilitating the *identification of causal relationships*. Claims about imaging technologies allowing us to "see causation" in the sense of picking up causal information are frequently made either implicitly or explicitly[163] by biologists. This is a very important idea to try to understand since, if true, it means that some 4-dimensional visual representations can provide us with causal information that is not only less accessible but that may not even be present in other data formats. With regard to this function of images, we need to distinguish between static images, whether 2- or 3-dimensional, and moving images such as the videos produced in conjunction with live-cell imaging techniques. Moving images are widely claimed to provide more information than static images and often the sort of information that we can get from them is couched in causal terms. The obvious candidate for the extra information contained in moving images is the temporal dimension. However, as will be discussed later, temporal data is not absent in all static representations: serial representations of some object created at defined time intervals can, at least in theory, represent the same information.

The idea that causation is something that we either can or cannot perceive has a long history within philosophy (though the majority of writers, including Hume, have taken the negative position) and there is a more recent, though still substantial, psychological literature investigating the question of when and how we get the visual impression of causation. It has, of course, proven to be very difficult to come up with a satisfactory account of what causation *is*.

[163] Though explicit claims tend to be limited to oral communication, a fact that perhaps already suggests that this claim is not to be taken at face value.

Fortunately, this question need not be resolved in order to answer the question of whether imaging cells (or other biological objects) in real time allows us special access to causal information. I will claim that, whichever interpretation of causation we accept, whether we can get causal information from data has less to do with whether it is in the form of static or kinetic images – or even images at all – and more to do with the background information we have about possible or plausible causal mechanisms. I will argue that while we do very often get 'extra' information from imaging events in real time, we do not specifically get causal information. Furthermore, the additional information we get is not specifically due to the data being presented in the form of a 4-dimensional image, but is rather due to features of the experimental set-up such as the temporal resolution that can be achieved using different methods or the ability to track single objects (cells or molecules) over time.

5.3. Why images? Some other perspectives.

Before turning to an examination of these two potential epistemic roles of images, I want to very briefly acknowledge some of the answers that other disciplines have to offer to the question of why images are a preferred form of biological evidence, even when other options are available. A considerable amount of work has been done on the history, sociology, anthropology, and rhetoric of scientific images (e.g. Dumit 2004; Lynch and Woolgar 1988; Cartwright 1995; Brain and Wise, 1994; Jones and Galison 1998; Elkins 1999; Kevles 1997; Abraham 2003; Breidbach 2002) and these perspectives are crucial to a complete answer to this question. While it is clearly impossible for me to address all of the responses that these disciplines have suggested, I want to very briefly sketch a couple of possibilities that seem to be particularly important. These are: 1) the historical importance of visual evidence and images in

medicine and biology, and 2) our affinity for and attraction to images together with the rhetorical power of images.

5.3.1. Historical Preferences

N.J. Berrill has claimed that biology is and has always been an "eminently and inherently visual" science (1984, 4). Evelyn Fox Keller claims that, while various branches of biology take different forms of evidence to be explanatory and there has often been conflict between those, in the tradition of natural history, that give preference to observation (whether direct or via imaging technologies) and those that are more theoretical and give preference to mathematical models, there is a common attraction to the use of visual representations that resemble what we get by direct observation - i.e. naturalistic images (2002, 202). Data that is the output of mathematical models – cellular automata, for example – becomes more acceptable to a broad range of biologists and gains persuasive power when the results are displayed in ways that bear visual resemblance to the objects and processes they are supposed to represent (2002, 272). Essentially, it seems, most biologists like to watch natural objects doing things. Advances in biological imaging, including confocal microscopy, in the last 15 years or so are widely considered to have revolutionized[164] cell biology. While this claim is true simply in virtue of the enormous advances that have been made in the types of questions that can be asked and the ease with which they can be addressed, it is often justified at least in part by making reference to the fact that these advances have allowed us to watch events occurring inside cells. It is not only that we now have the ability to easily ask many questions that were previously difficult or impossible to address: it is that we can see – or watch – things happen. I will have more to say about the difference between seeing and watching later, but for now it is sufficient to notice that

[164] This term, or "revolution" is used in almost every paper that makes reference to the period that is usually taken to begin with the discovery and cloning of green fluorescent protein (GFP).

having specifically visual access to objects and events of interest is a longstanding desire in biology.

Some forms of biological data may not obviously present alternatives in terms of preferred data display format: with confocal microscopy, for example, you see an image if you look at the specimen through the eyepieces or on the monitor so it seems natural that the output should be an image. In the case of PET, however, the choice seems less obvious since the data exists as numbers before it gets converted into an image. However, just as current cell biology may share some preferences for visual observation with the natural history of past centuries, PET also belongs to a lineage of technologies with a particular preferred format of representation. We can trace a direct line from the X-ray to CT (computed tomography) to PET.[165] Originally, X-rays were necessarily photograph-like, being produced by the direct interaction of X-rays with the film.[166] No mathematical processing goes into the production of an X-ray and while it can be measured after its production (e.g. by densitometry) and the data presented in some other format, the X-ray is essentially pictorial. CT is essentially a tomographic, 3-D X-ray format that requires image reconstruction to regain spatial information just as PET does. In the very early days of the technology the representational format made explicit acknowledgement of the mathematized nature of CT. Instead of a picture with each pixel assigned a color or shade of grey, the data was displayed as a two-dimensional array of numbers. The number indicating the intensity of each pixel was displayed in an arrangement

[165] For a history of medical imaging, see Kevles 1997.
[166] This is no longer true, however, since X-rays data is often collected digitally rather than by using photographic film.

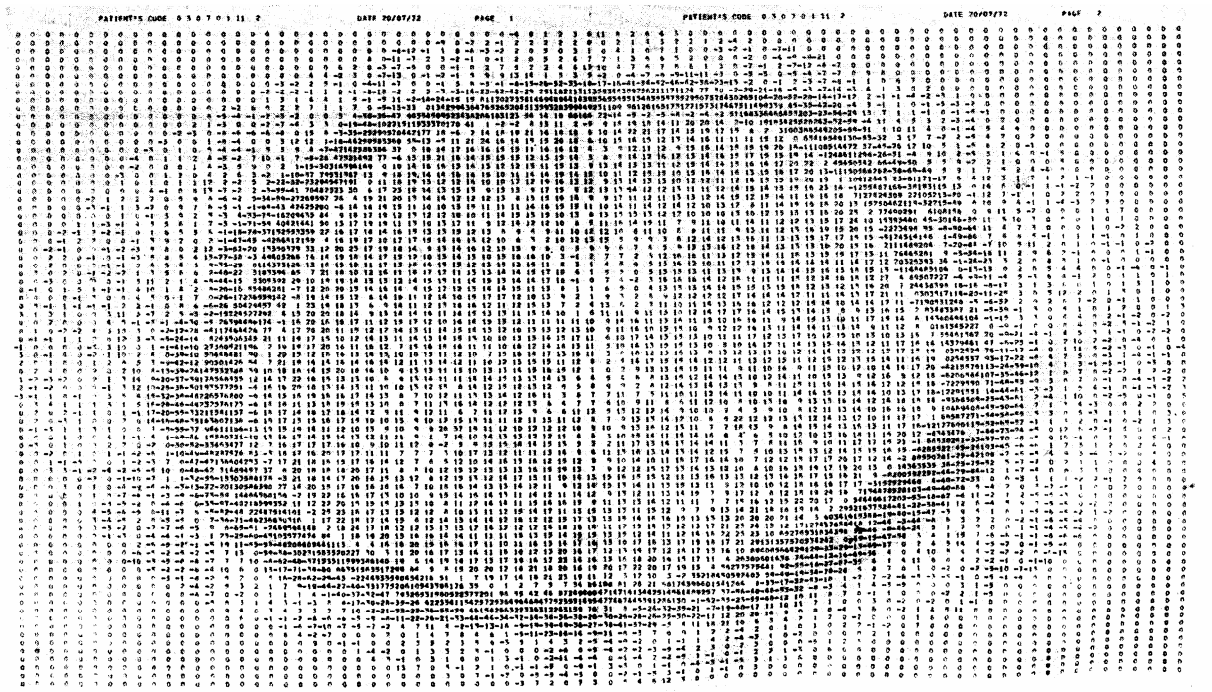


Figure 5.4 An early example of CT data.

CT data as shown in the initial clinical report on the technique. This figure appeared in Ambrose (1973).

approximating that of the brain (the early use of CT was for brain imaging) as shown in Figure 5.3. Very soon however, CT data began to be displayed in a standard pictorial format, like an X-ray. This might reflect the preferred representational style of the researchers and clinicians using the CT images[167] as well as the undoubtedly greater ease of gathering information from the straightforward image. So it might be that the naturalistic style of PET images reflects the pictorial preferences of a discipline in addition to (or rather than) conferring any particular epistemic advantage.

5.3.2. Affinity for and rhetorical power of images

The source of the preference of biologists for visual access to the world that was discussed in the last section is not entirely clear. One plausible way of accounting for it is by reference to the

[167] Along similar lines, in discussing the representational style of electron micrographs, Rasmussen (1997) claims that it was strongly influenced by the way that previous types of cytological images were presented.

fact that, when used under appropriate conditions, visual perception is usually reliable. We learn to trust the results of our eyes, under most conditions, and ways of investigating the world that seem to be like straightforward, unaided visual observation may more easily be taken to also be trustworthy in virtue of this apparent similarity. In essence, seeing is believing and if we can come up with new ways of seeing then we might at least be inclined to think that we should believe what we see in these new ways too. While of course no scientist naively believes that our eyes or imaging technologies always produce veridical data, the phrase "seeing is believing" appears in several paper titles[168] as well as in a recent letter to the editors of *Nature* in which the author suggests that our natural tendency to go from seeing to believing is now being inverted through the use of digital manipulation of image data.[169]

The editorial on which the author of this letter is commenting, brings up another reason why images may have persuasive power: we are simply drawn to attractive images. We like to look at them and we like to make them: "Tweaking images is also seductive in a way that adjusting statistics is not, because of the natural human desire to create an aesthetically pleasing picture" (Pearson 2005, 953).[170] This sometimes leads us into questionable digital manipulation practices, but it also leads to such things as the calendars of extraordinarily beautiful scientific images that are often put out by companies such as Zeiss that make microscopes. The beauty of the images may, in some cases, be an end in itself, but it may also serve other purposes. In 2003, the American Academy for the Advancement of Science together with the journal *Science* organized the first annual Science and Engineering Visualization Challenge. The report on the outcome of the 2004 version clearly states that the contest was designed to foster "the ability to

[168] Hearst 1990, Orr-Weaver 1995, Monteith 1995, Herschman et al 2000.
[169] Greene 2005.
[170] The idea that we like to create pictures that resemble the world around us can be traced back as far as Aristotle's claim that humans have a natural tendency toward mimetic activity (Poetics xxx).

convey the essence and excitement of research in digitized images, color diagrams, and even multimedia presentations" since this increases public attraction to and understanding of science and since it "the general public that ultimately supports the global research enterprise … everybody benefits" (Supplee and Bradford 2004, 1903). Joseph Dumit, an anthropologist of science, suggests that images can do this in virtue of their ability to serve multiple purposes and hold several different meanings simultaneously. A single PET image can represent not only the actual blood flow in a slice of a specific individual's brain over a particular time period, but also the pattern of blood flow in some *type* of person (e.g. schizophrenics), the viability of PET as a research tool for certain disciplines and types of questions, and (perhaps most importantly for the public perception and support of science), the value and importance of research in neuroscience more generally (2004, 4).

As suggested by the multiplicity of meanings and roles they can play, images are important not only for the reception of science by the general public, but for the evaluation of individual pieces of research and research projects by journal editors and grant review boards. Dumit interviewed a number of prominent PET researchers about various aspects of their use of images and found that most claimed that it was crucial to include brain images (as opposed to only graphical or other statistical data) in articles submitted for publication or grant applications since the failure to do so significantly reduces your chance of getting your work published or funded (2004, **57**). It is important to note that the quality of the data doesn't change between these different display formats, but apparently the appeal, power, or apparent importance of the data does. One feature of naturalistic images that potentially contributes to the authority that they may hold outside of a very specific scientific or medical context is their resemblance to photographs. Despite the enormous complexity of producing a PET image, by or for the

layperson such images are often interpreted as being essentially photographs of the brain. As such, they inherit the presumed objectivity and reliability of a photograph[171] and serve as persuasive evidence for the (multiple) claims that they are used to support. The combination of presumed objectivity and reliability together with the ease with which these images come to hold multiple meanings gives them enormous power. Others have written extensively about the power that visual images (scientific and otherwise) exert on public discourse and emotion (e.g. Mitchell 1994; Cartwright 1995) as well as on the aesthetics of scientific images (e.g Stafford 1991, 1994, 1996; Elkins 1999), but I will not be able to discuss their work here. Instead, I will now turn to the first of the two potential epistemic roles of images: cognitive accessibility.

5.4. Cognitive accessibility

The fact that visual representations (including not just photograph-like images but diagrams, maps, graphs, etc.) can present us with large amounts of complex data in a way that is more easily available to our cognitive apparatus than is data in a straightforward numerical format is uncontroversial. Even very simple types of visual array such as arranging numbers from highest to lowest makes it much easier for most people to identify certain features of the data (Tufte 1983, 1997). Faced with tens of thousands of numbers in a PET data set listed in a linear sequence proceeding from the first to the last slice (along the *z*-axis) and, within a slice, from left to right (along the *x*-axis) and from bottom to top (along the *y*-axis), no human could hope to identify regions of higher or lower activity within any reasonable period of time. We might be able to scan the list and eventually come up with a set of the highest numbers, but to keep track of the position on the *x*,*y*,*z*-axes each value belonged and which were adjacent to other high values, while theoretically possible, would be enormously difficult and time-consuming,

[171] Though awareness of the extent to which photographs are digitally manipulated has undoubtedly reduced the degree to which photographs are seen as reliable and objective representations of the world.

especially if we were not to add any sort of visual representation (lines, symbols, etc.) to mark the spatial location to which each number belonged. The epistemic value of cognitive accessibility, then, is not that images contain spatial information that is not present in the corresponding numerical data, but that they make it much easier to get it into our heads; to produce belief or knowledge. In general, the larger and more complex the data set, the greater the epistemic advantages of using some form of visual representation.

However, this advantage holds in general for any type of visual representation. Is there any special advantage to photograph-like images compared to other visual formats such as graphs? Recall that some scientists who use PET are reported to prefer graphs to semi-naturalistic brain images. An important caveat that was left out earlier, however, is helpful in identifying what advantage images specifically might have. Even scientists who prefer to analyze their data using various graphical representations use the images at earlier stages in order to get an overall sense of the data in order to judge whether the experiment worked or showed some characteristic(s) that might suggest that something had gone wrong with the experiment.[172] What the images do very effectively is to give the user a sense of the overall characteristics of the data: how both adjacent and distant parts of the image compare to one another and what both the global and local characteristics of the data are. Looking at the image, for instance, makes it easy to see whether two regions of interest (ROIs) are active at the same time or if one region becomes active following the other. The same information is present if the data is presented in the form of two graphs, a time course of activation for each of the two ROIs, but in this case additional work must be done (e.g. using the same scale and aligning one graph above the other) in order to pick out this larger scale feature of the data. Under some circumstances, looking at the image may

[172] Julie Fiez (Associate Professor, Departments of Psychology and Neuroscience, University of Pittsburgh), personal communication.

help to initially identify a ROI. While an ROI may sometimes be defined prior to an imaging study in terms of anatomical structure or Talairach coordinates, in other cases the ROI(s) may not be identified until the imaging data has been acquired. In such a case, an ROI will often be defined as an area that is differentially active between control and test individuals or between a baseline state and a response to some stimulus. Delineating the boundaries of an ROI in this situation involves identification of an area of high activity and so requires a very simple kind of "seeing as". Although the area may have no pre-specified shape, it still requires that we recognize specific features as characteristic of an ROI: a patch of either uniform color (or a mixture of the colors representing the highest activity levels) and the boundaries at which the color shifts to one representing a lower activity.[173]

This advantage that images have over representation of portions of the data in other formats such as graphs may sometimes be reversed, however. The central point about cognitive accessibility is that some types of representation allow us to more easily make certain sorts of discrimination. Sometimes it is easier to discriminate more local features of the data if the extra, non-local data is removed from view. The presence of excess information can make it harder (though, again, not impossible) to pick out specific features of interest. Thus, if what you are most interested in is relatively subtle differences in the timing of activation of a specific ROI between two populations, it may very well be easier to make the relevant discriminations by looking at time courses for that ROI in the two groups and eliminating all of the data from other areas. The effect of using different color schemes in Figure 5.2 is essentially the same: it highlights some differences while obscuring others and so facilitates some discriminatory tasks while making other more difficult.

[173] The same information could be extracted from the numerical data but it would almost certainly require using a computer and some sort of pattern recognition tool to identify the required features. Except perhaps in very simple cases, we could not easily identify these features from numerical PET data on our own.

Representation of the data in numerical form can also offer some potential advantages. In particular, numerical data is capable of representing an infinite number of different values. Of course, no instrument offers infinite precision so that the full advantage of numerical representations is never actually needed. But a numerical format does have the capacity to fully capture the granularity of the instrument: the apparent granularity of the numerical representation will be equal to that of the instrument whereas the apparent granularity of the data represented as an image will usually be less than that of the instrument. This situation arises because the human visual system is able to discriminate only a very limited number of shades of grey and larger, but still finite, number of colors. Therefore, when PET or other data is represented as a grey scale or pseudocolor image, each grey level or color must be used to represent a range of numerical values. This is not necessarily a bad thing – in fact, by eliminating some differences that are not relevant to answering the question of interest, we can more easily pick out those that are relevant.[174] However, it does mean that finer-grained distinctions can be made over the full range of data values using the numerical data rather than an image. The full granularity of the instrument could be captured in a set of images if we iteratively selected small regions of a larger scale image in which only a limited segment of the full color range was present and redefined the intensity range associated with each color such that a smaller and smaller range was used in each iteration. We could eventually capture the full granularity of the instrument, but it would require a large number of images and the fact that the same color would represent different intensities in different images would eliminate any cognitive advantage.

[174] Trying to minimize the range of values represented by each color by using as many colors as the human visual system can discriminate would only reduce the ease with which we could make any discriminations.

Thus, the cognitive advantage of images over other forms of data display is relative. The existence and extent of the advantage is dependent of the sort of discriminations that need to be made and, even in cases where there is a significant advantage to using an image, the image is easier to use but does not allow anything to be done that could not in principle be accomplished using other forms of data.

5.5. Perception of causation

Do images have an epistemic advantage over other types of data display in that they, and not other formats, give us access to causal information? More specifically, do moving images – videos – created by imaging living cells allow us to pick up information that we cannot get either from either series of static images or other forms of data? Since numerical data precedes both the static images and videos, there need not be any difference in content between any of these forms of data display.[175] This is important since it means that we can isolate epistemic differences that are due to the data display format from those that must be attributed to different object-representation relationships. It is differences that can be attributed to differences in display format that are the primary concern of this section.

The advantages of live cell imaging are widely celebrated though it is often a bit unclear precisely what the advantage is (all italics my own):

> "Static images – until now the source of most data in developmental biology – give an *incomplete* view. … [Live cell] imaging allows scientists to take advantage of the world's fastest computer processors: their own eyes and brains. Humans can take in lots of visual information at once and extract *patterns* from it;

[175] As will be discussed shortly, though it is in principle possible for numerical data, static images, and videos to have the same content, in practice the experimental set-up and data are usually different when one compares static images to videos. These differences are very important but they are not differences that are due to the data display format itself.

> complex images and movies provide such information."(Beckman 2003, 76)
>
> "With the advances in labeling and imaging technologies, we have already witnessed remarkable improvements in our ability to monitor and *interpret* processes in real life and in real time." (Hurtley and Helmuth 2003, 75)
>
> "Being able to observe processes as they happen within the cell by light microscopy adds a vital *extra dimension* to our understanding of cell function." (Stephens and Allan 2003, 82)
>
> "The ability to visualize, track, and quantify molecules and events in living cells with high spatial and temporal resolution is essential for *understanding* biological systems. […] the development of highly visible and minimally perturbing fluorescent proteins … together with updated fluorescent imaging techniques, are providing unparalleled insights into the movement of proteins and their interactions with cellular components in living cells" (Lippincott-Schwartz and Patterson 2003, 87)

The above quotations were all taken from articles in a special section of *Science* devoted to biological imaging. Live cell imaging is identified as giving us more (complete, with an extra dimension) and better information (allowing extraction of patterns, interpretation, understanding), but what is the nature of this extra information? The obvious response is that it gives us temporal information, but, as noted earlier, temporal information is not absent from all data presented as static images. Series of static images produced at defined temporal intervals also convey this sort of information, though it is often less fine-grained temporal information and is added to the images rather than being strictly contained by them. However, coarser temporal resolution is a matter of different content of the data rather than a necessary feature of the data display format and time lapse video data also requires that some temporal information be added

back to the video images, so these differences will not play a role in identifying possible epistemic differences specific to data format.

Is it only the case that live cell imaging usually gives us more of the same sort of information that is present in static images, or can we get a different kind of information from moving images? Though none of the review papers cited above make any direct reference to causal information, if we look at reports of specific imaging studies, they tend to report primarily two types of information: 1) descriptions of the spatiotemporal movements of one or more objects, and/or 2) calculation of the dynamic or kinetic features of specific interactions. It is the first sort of information that is most relevant here since the descriptions involved usually include ascriptions of causal relationships between various imaged components. Thus, for instance, we read (italics mine):

> "…that kinetochores can attach to the forming spindle by *capturing* astral MTs [microtubules] was *directly demonstrated* by video microscopy … Subsequent video microscopy studies revealed that this kinetochore switches between two activity states: one that allows it to *move poleward in response to a force*, and another that allows it to be *pushed (or pulled) away*" (Rieder and Khodjakov 2003, 93)

> "Growth of phragmoplast across the cell creates a new partition in its wake, giving the *visual effect of a curtain being pulled across the cell*. Throughout this process, the advancing front of the phragmoplast is in intimate contact with the parental wall, suggesting that *short-range interactions between the phragmoplast and plasma membrane may play important roles in guiding the cell plate* throughout much of its development."(Cutler and Ehrhardt 2002, 2812)

Such causal claims are, as above, often explicitly based on the interactions that are seen in the videos. Is it the case, then, that watching videos produced by imaging living cells allow us to identify causal relationships in a way that we cannot by seeing static images or numerical data?

Causal information is central to scientific explanation, so the identification of causal interactions is crucial to our understanding of the various objects and events studied in biology and medicine. We want our methods to allow us to do more than simply describe what happened: we want to understand how and why things occur. The question of whether we can get such causal information specifically by watching is considerably more complex than the issue of cognitive accessibility for three reasons. First, the concept of causation has proven to be notoriously difficult to pin down. If we want to understand whether we can "see causation", we need to have some idea of what makes a sequence of events a causal sequence and of how the relation of causation is identified or defined. In particular, we would like to know whether the causal relation is something that supervenes on other physical properties of the world and what features of the world, if any, we need to observe. Second, the idea that 4-dimensional images in particular might provide us with information about causal relations is often found in the context of a comparison between data obtained in different ways – for example, between samples that were taken at various times, then prepared for imaging and a single specimen that was monitored continuously or near-continuously over an extended period of time. Moreover, the techniques involved are not techniques merely for visualizing, but involve intervening in the systems that they make visible. Thus the question of whether we are able to extract different information from different ways of displaying the same data needs to be separated from the question of whether some methods of intervention provide more, less, or different information. Third, we need to address the question of how the content of the data and/or data display is related to the psychological effect (or causal impression) that seeing certain sorts of visual interactions produces in humans. The two are not always connected. We may, for instance, get a visual impression of causation from only a subset of the interactions that represent actual causal

relations.[176] Alternatively, we may get an impression of causation from visual interactions that are not causal.

Each of these three features will need to be addressed in turn. The second is more easily separable from the first and third, so I will begin with it.

5.5.1. Different data

As it was in the quotations in the previous section, a contrast is usually drawn between the information we are able to get from seeing static images and from watching videos of living cells. Static images, even when they form a time series, are taken to be incomplete relative to video and to be unable to provide the information we need to fully interpret and understand protein and other molecular interactions. The question at issue here, though, is not whether videos simply contain *more* information than static images, but whether the 4-dimensional data display format provides us with a specific kind of additional information – a kind that allows us to extract causal information about the objects and interactions that are imaged. In order to do this, we need to compare apples to apples. The problem with comparing series of static images to video is that they usually involve not only different forms of data display, but different ways of intervening with the objects of investigation. They are different in at least three important ways. First, video will normally allow a much greater temporal resolution. Second, sampling requires that each image in the time series represent different individual cells/molecules while live cell imaging can continuously monitor a single cell or molecule from start to finish of the imaged process.[177] Third, for many objects and events there may exist no way of monitoring them or making them visible by any technique other than those normally used for live cell imaging.

[176] The causal anti-realist should replace "actual causal relations" with something like "interactions of the sort that actually meet the criteria for what we call causal relationships".
[177] Keller also refers to the first two differences, but not the last, in her assessment of the advantages of studying living cells (2002, 225).

With respect to the issue of temporal resolution, while neither human perception nor any type of video are truly continuous,[178] they can much more closely approximate continuous monitoring of some object than can any technique that depends on taking samples at discrete intervals. The sampling interval that is technically feasible in generating a time series will depend in large part on the nature of the particular experiment. If, for instance, I want to examine the movement of a particular protein in a cell undergoing mitosis, I can synchronize a large number of cells in a flask, extract samples at regular intervals (maybe every 5 seconds, if I have help), prepare each sample (*e.g.* by fixing then staining the cell with a fluorescently-labelled antibody), then photograph them using an appropriate kind of microscope and camera. The sampling interval will be determined by how quickly I can withdraw samples as well as by how quickly I can halt the process in which I'm interested. With an automated system, I might be able to get below 5 seconds, but there is no way in which I could achieve a sampling interval of 1/300 second to correspond to the frame capture even of a CCD camera. As a result, I am virtually certain to get data with a much finer temporal resolution by using my confocal microscope with a living cell (or group of cells) in which the protein of interest has been made visible in some way (perhaps by creating a GFP fusion). By keeping the cell(s) alive in culture for whatever time period mitosis requires in that species, I can get the same spatial resolution as in the previous case, but much greater temporal resolution.

The difference in temporal resolution alone may or may not make an important difference to our ability to make causal claims based on the video vs. the time series data. Whether it does or not depends on the temporal scale at which events relevant to ruling out the occurrence of alternative interactions or mechanisms, as will be discussed later. The important point, though,

[178] Our visual system involves brief saccades as well as attentional shifts that prohibit continuous attention being paid to any single object. CCD cameras can now capture frames at about 300 frames per second.

is that this difference is not one that is due to differences in content between data display formats, but due to the fact that the video format almost always contains more information (but of the same sort) than does the time series. To show that there is a difference specifically due to the type of data display would require that we be able to get additional information from the video than from a series of static images spaced at intervals equivalent to the video capture rate (such a series could be created from the video). This, however, is not the situation that is referred to in the above quotations or in any discussion of the advantages offered by live cell imaging. From the perspective of the scientist, this comparison is of little interest since time series data with the same content as video doesn't actually exist. What matters is simply that you can ask and answer more questions using live cell imaging methods. For the philosopher seeking to understand the epistemic significance of watching, however, it is very important to distinguish differences that are due to the amount of data that is collected from those that may be due to the difference in data display format.

The second contributing factor to the difference in the content of data in video format compared to series of static images is the visualization of a single individual vs. multiple individuals of the same type. This difference is impossible to eliminate, even in principle, since taking and preparing samples at timed intervals is a destructive process. A cell that has been removed from culture, fixed, and stained at time *t* is of course no longer available to be sampled at time *t+1*. Different samples must inevitably be used at each time point.[179] How significant this difference is will depend on the range of variation between individuals – the greater the variation, the greater the impact. However, it is not the case that live cell imaging is entirely free

[179] The sample should not be confused with the entity (organism, types of cell, etc.) from which it is being drawn. For example, we can take multiple biopsies of the same tumor at different times, but the cells in each biopsy are different.

of this concern with the effect of differences between individuals.[180] If the object of interest is whole cells (or larger units), then it is almost always possible to monitor individuals. But a large part of the advantage of techniques such as confocal microscopy of living cells is not the ability to image larger units such as cells, but to monitor the spatiotemporal activities of specific proteins. While it is increasingly possible to visualize single molecules and to follow an individual molecule through some event,[181] it is still more common that many molecules are labeled and that all of the molecules present in a given cell or cell compartment cannot be distinguished.[182] In such cases, it is aggregate behavior of the labeled molecules that is observed. Accordingly, unless single molecules are imaged, there is only a difference in degree between the effect of a time series generated using different individuals and a video in which the change in spatiotemporal position of multiple molecules is represented (for instance, as a general shift in fluorescence intensity from the cytoplasm to the plasma membrane)..

The third difference is that technical limitations with other methods mean that live cell imaging may sometimes provide the only means of addressing certain questions. In this case, the fact that the video display format contains more information follows simply from the inability to effectively intervene using other methods whose usual output is static images. There may, for instance, be no good antibody with which to stain some particular protein in fixed specimens. GFP and its variants, however, can be genetically fused to virtually any protein of interest, usually without interfering with the function of the protein. Even if alternatives are present, creating a GFP fusion and using the fluorescent protein to track or quantify single or multiple

[180] This is not to deny that knowing the range of variation in a sample or population is a good thing, but only to indicate that knowing individual rather than aggregate behavior is also of value.
[181] See, for instance, Seisenberger et al. 2001; Murakoshi et al. 2004; Lommerse et al 2005; Ritchie et at. 2005; Koyama-Honda et al. 2005.
[182] In the case of GFP fusion proteins, for instance, every protein expressed from the altered gene will contain the GFP moiety. Additionally, new copies of the tagged protein will continue to be produced as long as the gene continues to be transcribed and translated. These new proteins are not yet in their eventual sub-cellular location, nor are they contributing to specific functions.

proteins in the cell is often much faster and easier than taking and preparing individual specimens from multiple time points. This is without any question an enormous advantage for live cell imaging techniques. However, it is an advantage that belongs to the intervention aspects of these methods rather than the visual display aspects. Moreover, this difference could easily be overcome by simply creating the GFP fusion and using cells containing it to generate a time series of static images. No one in their right mind would do this since live cell imaging requires far less time and effort, but in theory the difference could be eliminated.

To sum up this section, then, there are indisputably many advantages associated with live cell imaging. However, many of these follow from the fact that imaging technologies are tools not only for visualization but for intervention. Differences in data display format (video vs. static images) usually also involve differences in the type of manipulations that have been performed on the biological system. While the differences that are specific to the intervention aspects of the methods are very important for scientific practice, for the purpose of this chapter they need to be separated from the epistemic effects of different forms of data display. In order to assess the effect that the data format may have on our ability to get causal information, we need to start with the same data. The general claim that videos may allow us to get causal information about protein interactions while times series do not requires that we compare the video with the same data presented as a time series (i.e. a series of static images consisting of each frame captured in the video). Specific causal claims will require different amounts and types of information and, in some cases, video microscopy may produce far more data than is required and a series of static images at, for instance, much lower temporal resolution, may contain all the necessary information. Further discussion of when and how this might occur will follow in the next section.

5.5.2. Understanding causation

In order to determine whether or not we can get causal information from some data display formats and not others, we first need to have some idea of what we mean when we say that we *saw* X knock over Y or make some other sort of causal claim. Hume famously argued that causes are not knowable *a priori* and that the observation of regularities in the world serves as the basis for our impression of causality. However, all we really perceive, according to Hume are constant conjunctions of objects or events. When we see one event regularly followed in space and time by another, the first one will naturally and forcefully bring to mind the expectation of the second.[183] The causal impression just is this action of the mind. Although one can see the prior event that would be labeled 'cause' and the subsequent event that would be labeled the 'effect', it is not possible to see a causal connection between them. Other philosophers have contested this claim, maintaining that it is indeed possible to observe causation, even in single cases where no constant conjunction can be found. Thus, for instance, Ducasse (1926) contends that by observing the relata of a causal relation, we observe the cause. Anscombe claims instead that the concept of a causal relation is too abstract and has meaning only if we can first understand ideas like push, pull, break, bend, etc. (1975). These causal concepts can be applied on the basis of observation – we know what it looks like for something to break or to push or pull another object. What we see, then, are instances of pushing, pulling, breaking, etc, not causation more generally.

There is a very large philosophical literature on this topic[184] and I cannot hope even to review all of the many accounts that have been developed let alone resolve the question of what

[183] As many have pointed out, Hume actually seems to have two different definitions of "cause" - one based on regularity and the other making reference to counterfactuals.

[184] Not to mention a great deal of dissatisfaction with the state of the discussion: "The attempt to 'analyze' causation seems to have reached an impasse; the proposals on hand seem so widely divergent that one wonders whether they are all analyses of one and the same concept." (Kim 1995, 112).

causation is. However, two very general sorts of alternatives can be identified. The first is Humean and holds that what we are doing is drawing an immediate, automatic inference based on background knowledge of some sort. Exactly what this background knowledge is supplying that allows us to identify certain relations as causal – information about regularities, counterfactual dependence, or something else – is far from being agreed upon. But for the present purpose, disagreements of this sort can be passed over since the crucial feature of this position is that it holds that we are making an inference based on background knowledge of some sort, though we make the inference automatically and unreflectively. The second alternative is non-Humean and claims that causal relations are immediately accessible to experience: we really are *seeing* causation, just as we would see a color or a shape. Anscombe (1971) and Ducasse (1926) offer the most prominent defenses of this sort of view. The former alternative is easily the dominant one, but the latter is not without its current defenders. Interestingly, many psychologists working on visual perception claim that, despite what Hume claimed, we can perceive causation. Thus, it will be informative to consider some of the arguments and experiments that psychologists offer in support of this view.

While it may be an open question which of the two alternatives more adequately describes what is happening when we make causal claims on the basis of seeing some interaction at the macro level, I will argue that the Humean interpretation is the only possible alternative at the micro level of confocal and other types of microscopy since the careful application of background knowledge is indispensable to making causal attributions based either on seeing or watching.

Minimal conditions for seeing something as causal

What is needed for my argument is not a complete theory of what causation is but an account of the minimal conditions that are required for us to be able to see that or how X caused Y based on the information available in some type of data display. These conditions, I claim, are as follows:

1. In order for there to be a causal relationship between two objects or events, they must possess spatiotemporal contiguity with certain acceptable limits.
2. Importantly, these limits are determined not by characteristics of the data alone, but are established by background information regarding the sorts of interactions that are possible or impossible for objects or events of the involved type(s).

The reasons for identifying these particular conditions will be elaborated in the remainder of this section. Two features are important to notice at this point: first, spatiotemporal information can be obtained from any data format, so there is no special ability to get causal information from video displays, and, second, the role of background information is what favors the Humean alternative, at least at the micro level.

If it were to be true that we can get causal information in videos that we cannot get from static images or other (numerical, statistical) forms of data display, causal information would have to follow either from some difference in the content of these displays or from our psychological response to them. Since there is no necessary difference in content between different data display formats, as has been argued earlier, I will turn to the question of the psychological effect.

Psychologists on perceiving causation

Since the early work of the French psychologist, Albert Michotte, in the middle of the twentieth century, a large amount of work has been undertaken to investigate when the visual system will interpret a dynamic stimulus as causal (e.g. Michotte 1946/1963; Scholl and

Nakayama 2002, 2004; Twardy and Bingham 2002; White and Milne 1999, 2003). This work involves experiments such as the following (see Figure 5.4): two shapes, A and B, are displayed and animated on a computer screen. A begins to move towards B then stops when it is immediately adjacent to B. Just when A stops, B begins to move away from A.[185] Observers

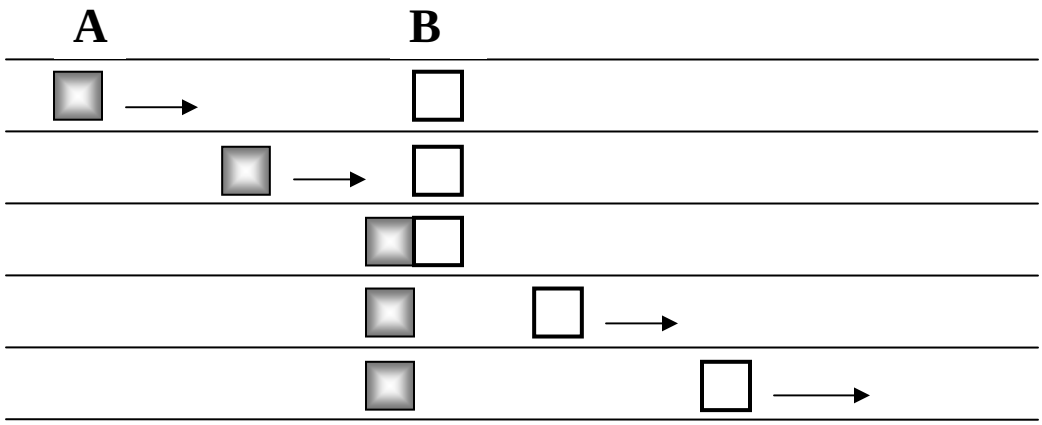


Figure 5.5 Illustration of Michotte's launching effect.
Adapted from Thinès, Costall and Butterworth (1991, 69)

(adults or children) are asked whether or not A was the cause of B's motion. In the situation just described, the majority of observers will claim that A was the cause. However, if the set-up is changed slightly so that B starts moving before or after A stops[186] or if A overlaps B before B starts moving, then the proportion of people who claim that A caused B to move drops significantly. Michotte referred to the "illusion" of causality in this interaction as the "launching effect". Other sorts of interactions have also been shown to produce the impression of causation:

[185] There are additional factors that matter, most importantly the relative speed of A and B.
[186] A delay of up to about 50 ms between the collision and the departure of B did not affect whether or not people reported seeing the event as a "launch". Gaps of up to about 100 ms reduced the proportion of people who claimed to see the interaction as causal while gaps of more than 150-200 ms eliminated the effect.

entraining (when, after colliding, A and B move off together in the same direction as A was moving before the collision), pulling, enforced disintegration, bursting, and penetration.[187]

The most important thing to notice about these experiments is that they indicate that there are tight constraints on the sorts of interactions, whether viewed once or repeatedly, that humans perceive as causal. Some visual interactions almost always, seemingly unavoidably,[188] produce the impression that they are causal, while others, no matter how regularly they occur, never do. This suggests that we must reject the claim that scientists can directly "see causation" in videos of living cells. The animations used in these experiments are very simple, even when contextual factors are added in to see how they influence the perception of causation. Cases of observed interaction in, for instance, a confocal microscope are almost always much more complex than the simple interaction just described. There are many more objects and the types of interactions will not often fall neatly into one of the above categories. Thus, the visual interactons observed in a confocal movie will almost certainly not produce a causal impression in the sense described by Michotte and others. Moreover, even if such a straightforward type of interaction were to be observed, there is no reason to think that the causal impression corresponds to any actual causal relation. After all, no actual causation is involved in the animations used by psychologists. We may "see" causation where it fails to exist and fail to "see" it where it is present. Many cases where we may have good reason to say that a protein-protein interaction involves some sort of causal relation, for instance, will not involve the sorts of visual interactions that people identify

[187] Scholl and Nakayama 2004.

[188] I do not mean by this that it is innate, but only that we do it automatically and apparently without reflection. The question of whether it this is an innate or learned tendency remains a matter of considerable debate among psychologists who work on event perception. Experiments showing that 6 month old infants seem to perceive causation are sometimes taken to suggest that it is innate, but since even 6 month old babies have considerable experience with the world this conclusion is open to question.

as causal (e,g. launches) and , even if they do, will not fall within the correct spatiotemporal boundaries

Thus, the conclusion that we cannot "see causation" in this simple sense in 4-dimensional images does not mean that we cannot get causal information from a movie (or from other forms of data). In order to clarify how we might do so, it is necessary first to look at the sort of information that seems to be involved when observers see certain interactions as causal. This will also help with elucidating my minimal conditions for causation. I have suggested that the information that we need to get from the data itself is spatiotemporal relationships between objects. But we also need additional information to interpret (rather than simply "see" in the above sense) an interaction as causal whether or not it is also "seen" as causal. This information, is background knowledge of the types of mechanisms that are plausible[189] in a given context.

What are we seeing when we "see causation"?

It is not clear exactly what is meant by the term "causal" in the animation experiments described above. Michotte simply asked observers how "causal" an interaction seemed to be or, alternatively, to give a free response describing the interaction (e.g. A pushed B, A crashed into B and made it roll away, etc.). What seem to be involved, however, are certain types of spatiotemporal relationship. A launch event is perceived, for instance, if a moving object gets close (enough) to another then stops, and the second object, after a suitably small time interval, begins to move in a certain direction with a suitable velocity. What counts as suitable in these instances is, presumably, determined by some part of the human visual system or other cognitive apparatus. Suitability may also be determined by background information in cases where we *interpret* an interaction to be causal, as will be discussed shortly, but for the moment we can

[189] Or at the very least, not impossible.

ignore those modifiers and focus on the fact that the basic features that seem to be relevant are the relative spatiotemporal positions and velocities of the two objects.

Twardy (2000) similarly wants to identify the physical quantities that we pick up on in observing causation. He, however, adopts a Salmon-Dowe (e.g. Salmon 1984; Dowe 2000) conserved quantities account of causation and contends that what we really pick up on are the transfer of conserved quantities such as energy and momentum. While my project is very different from his and I want to avoid defending a particular theoretical account of causation, I do not think that Twardy's account helps with biological imaging. Cells, proteins, and other biological entities are not immune to physical laws, of course, but information about relevant transfers of conserved quantities are not usually going to be available via biological imaging methods. If, for instance, protein A phosphorylates another protein, B, which, once phosphorylated, undergoes a conformation change and dissociates from some third protein, C, all we will likely see (depending on the specific imaging technology used) is that A made contact with B and then B moved away from C. These are changes in spatiotemporal position, not energy, momentum, mass, or some other conserved quantity.

However, we cannot get sufficiently fine-grained spatiotemporal information to be able to recognize specific biological causal concepts like phosphorylation. We are unable to give precise descriptions of the spatiotemporal characteristics of many of the sorts of events or processes that we want to say are causally responsible for some change. What is involved, therefore, is not something like seeing pushing, pulling, breaking as Anscombe claims is at the root of our observation of causes. We don't know what it looks like for A to phosphorylate B in the same way that we know what it looks like for A to "launch" B or for one person to hit another. And even if we did have this knowledge, the resolution of most of our imaging

technologies is insufficient to allow us to discriminate between different causes on the basis of their appearance alone. An interaction between a GTP-binding protein and a GTPase activating protein (GAP) that causes the hydrolysis of GTP may well look just the same as kinase A phosphorylating protein B in a confocal movie. What we can observe is the changing spatiotemporal relationship between the two proteins and other parts of the cell, not the supposedly causal relation (hydrolysis or phosphorylation) itself. To determine which of these processes is actually occurring requires additional information about which proteins have been labeled and what sorts of activities they may engage in. This is not information that is present in the imaging data, whatever format is used to display it. The crucial role of background information in identifying causal interactions will be discussed shortly, but before turning to it the issue of whether spatiotemporal information of the required sort is only present in some forms of data display still needs to be resolved.

It should be obvious that information about the relative spatiotemporal positions of various objects is present not only in movies, but also at least in series of static images.[190] Spatial information is of course present in each static image and temporal information is present as long as the time interval between images in the series is known. As discussed earlier, differences in the temporal resolution of a series of static images and a movie are due to the intervention aspects of the imaging methods, not to the data display format. If we had a series of static images at time intervals equal to the inverse of the frame collection rate of the video or CCD camera, the two display formats would contain exactly the same spatiotemporal information. In practice, the temporal resolution for the time series will usually be much lower, but whether or

[190] The same information is present in numerical data. This information can be extracted; though with more difficulty, since the ability to recognize an object (in particular, its boundaries) is aided by visual presentation. However, one could use computational methods to identify objects (as they move and change over time) in the numerical data.

not this makes a difference to our ability to extract causal information from the data will again depend on the background information we have. Let us turn, then, to the role that background information plays in combination with spatiotemporal data, however displayed.

Background information

As I suggested above, background information is necessary in order to identify causal relations in any data display format: movie, static images, or numerical. In most cases of biological imaging, we will not get a Michotte style causal impression from a movie (and it is in principle impossible for us to do so from other data formats), but whether or not we "see" a causal relation in the data, we must have background information about the types of interactions that particular objects in particular sorts of spatiotemporal relationships to one another can, might, or cannot participate in. Background information supplying the possibility of there being a plausible mechanism for a causal relation would be required for any more than a descriptive statement about the spatiotemporal positions, and changes therein, of the objects under investigation. A launch, pull, penetration, or other interaction may be *seen* (in 4-dimensional format) as potentially causal, but can be *interpreted* as actually involving causation if there is potentially a mechanism that identifies the smaller scale causal concept (phosphorylation, etc.) and can so explain this interaction as causal.[191] The same sort of information is required if the interaction is *not* seen as causal. Given the severe constraints on the sorts of interactions that we "see" as causal, very few biological interactions will be "seen" as causal, but this has no impact on whether or not they can be interpreted as causal. Background information is crucial for supplementing spatiotemporal data, however the spatiotemporal data is displayed. In fact, it is often required even for observations made using unaided perception. If I see a clear liquid being

[191] While she is describing infants and children rather than scientists, Schlottmann (2001) describes prior knowledge of mechanisms as serving to constrain perceptual causality.

dropped onto a group of cells and see them disintegrate, I need to know what the composition of the clear liquid is in order to determine whether it was the cause of the disintegration. If it was a highly acidic or basic solution or a strong detergent, for example, I might reasonably conclude that it did cause the cells to disintegrate since my background information provides me with a causal concept (e.g. disruption of the lipid component of the cell membrane by detergent). If, however, it was a salt solution (with physiological solute concentrations), then it is very unlikely that it was the cause and I ought to look elsewhere for my explanation - perhaps there was a sonicator operating but I didn't notice because the volume on my iPod was turned up to a similarly cell-shattering volume.

One role for the background information is to identify the (possible) small scale causes actually involved in a larger scale (non-causal) interaction such as A approaching or moving away from B. Another is to rule out possible alternative explanations of some event, resulting in the possibility that an actual (low temporal resolution) time series of static images to be epistemically equivalent to a kinetic image in cases where the set of possible causal events or interactions occur at a time scale greater than the interval between images in the series. Just as there are constraints on the spatiotemporal conditions under which we will "see" a launch or other causal event, background information places upper and lower bounds on the larger scale spatial and temporal relations (those that we actually observe) that can be connected to the smaller scale causal interactions such as phosphorylation that need to be either ruled out or permitted. It is not necessary, for instance, that for one object to be claimed to have caused another's motion, that there be no temporal gap. If phosphorylation or some sort of conformational change is supposed to be initiated by the arrival of A close to B and responsible

for causing B to start to move away, it is entirely reasonable to expect that there will be a gap between the arrival of A and the departure of B.

5.6. Conclusion

The data that is acquired by many biological imaging technologies can be presented in different formats: as static images, as movies, as graphs, as diagrams, or even as very large sets of numbers. Images, however, are the dominant form in which the data is displayed. Why should this be? While historical, sociological, and rhetorical parts of the answer are important, the primary concern of this chapter has been with whether images confer any epistemic advantage over other formats. Two possibilities were raised: that the data is more cognitively accessible to us when it is presented as images, and that images – specifically 4-dimensional images – contain more and different information than other data formats. Of particular interest was whether 4-dimensional images permit us to get causal information that we can't get from static images or numerical data. The first possibility was found to have significant merit, especially for very large, complex data sets such as those obtained via PET and confocal microscopy. The second was found not to hold, at least not when the form of visual display is treated independently of other factors. While live cell imaging does often get us more information – sometimes causal information – than we can get from series of static images, this is due to the kinds of intervention that different imaging technologies allow rather than the data display format.

6. Conclusion

The starting point for this project was the question of how to understand the epistemic status of data produced by heavily mathematized imaging technologies such as PET and confocal microscopy. There are many kinds of mathematized imaging technologies and they play an increasingly important role in virtually all areas of biology and medicine. Some of these technologies have been widely celebrated as having revolutionized various fields of study while others have been the target of substantial skepticism and criticism. Thus, it is essential that we be able to assess these sorts of technologies as methods of providing evidence. They differ from each other in many respects, however one feature that they all have in common is the use of multiple layers of mathematical and statistical processing that contribute to data production. This feature alone means that these technologies do not fit neatly into traditional empiricist accounts of the relation between observation and evidence. Yet, in many cases, these instruments appear to live up to the claims of their supporters and provide very high quality evidence. Thus, it does not seem to be the case that their failure to fit into standard accounts of evidence reflects some general inadequacy on their part. In order to understand these technologies, then, we were led to look more closely at old philosophical questions concerning the role of experience and observation in acquiring beliefs and knowledge about the external world and saw that a more refined version of empiricism was needed in order to properly understand how these instruments can produce good evidence.

A number of relatively diverse positions have been labeled as “empiricist” over the last several hundred years. These range from the British empiricists of the seventeenth and eighteenth centuries to the logical empiricists of the early twentieth century to contemporary

versions such as van Fraassen's constructive empiricism (1980). In general, empiricism can be understood to encompass two closely related, but distinct theses. The first is a theory of meaning and holds that all of our concepts must be derived from experience. The second is a thesis about knowledge and holds that beliefs[192] must ultimately derive their justification from sense experience. These two doctrines do not necessarily entail one another, and it is the second one that has been central to more recent work in philosophy of science and is the one with which I have been concerned. The epistemological thesis can be interpreted in a stronger or a weaker sense. The weaker sense holds that we must use sense experience to make epistemic contact with the world; we cannot rely upon thought alone. The stronger sense of empiricism, however, is an anthropocentric one which holds that sense experience provides a uniquely high degree of epistemic warrant for our beliefs about the natural world, one that cannot be achieved by other means. Though the first, weaker sense is evident in many discussions of empiricism (e.g. Norton 2004), I have claimed that the anthropocentric sense either explicitly or implicitly underlies attempts in philosophy of science to provide an account of observation that extends the epistemic privilege of unaided human perception to other methods of data collection such as microscopes (van Frassen 1980; Shapere 1982; Hacking 1983).

Both senses are inadequate when it comes to understanding modern imaging technologies. The weaker sense, as it stood, insisted only that we need to observe the world in order to get knowledge about it. This usually means doing experiments and doing experiments of any kind[193] will involve using our senses. This is true whether we are using our unaided vision to watch squirrels cache and retrieve food or whether we are looking at a printout from fluorescence activated cell sorter that was used to determine the proportion of different cell types

[192] Here, I have been concerned specifically with beliefs about the natural world.
[193] Except for thought experiments (see Norton 2004).

in a sample. Accordingly, this sense failed to make any distinction between looking at some object of interest directly and looking at any sort of representation of that object – a picture of it generated by any means (photograph, painting, etc,), the color of the contents of a well in a microtiter plate whose intensity indicates the protein concentration of the sample placed in that well, the numerical printout that indicates the protein concentration of the samples in each well of the microtiter plate, etc. Thus, while not incorrect, in its original form, the weaker sense failed to be useful in trying to account for when and how modern scientific instruments provide good evidence.

The stronger, anthropocentric sense was also shown to be inadequate. The epistemic privilege associated with direct or unaided observation is based on the idea that evidence gathered using our unaided senses is supposed to be particularly reliable as long as we are using the sense in question[194] correctly and appropriately. There are differences of opinion regarding the type of aid to our native senses that can still be counted as observation (van Fraassen 1980; Shapere 1982; Hacking 1983). Claims about the limit or scope of observation use what I call *benchmark strategies* to establish the boundaries of observation by identifying instruments that bear a relevant sort of causal similarity to unaided human perception. Disputes arise because of disagreement about what sort of similarity is relevant. Chapter 2 argued that no existing account of observation contains a satisfactory notion of what the required sort of relevance is since none identifies an *epistemically* relevant similarity. Relevant similarity is supposed to be what justifies the claim that data gathered using certain instruments has the same epistemic status as data acquired using our unaided senses. Reliability (under appropriate use conditions) is what

[194] Usually it is vision that is at issue, though hearing, smelling, and touching may also be good sources of information in the appropriate context.

provides epistemic security, so relevant similarity must be defined in terms of the reliability-making features of human perception. This is what I referred to as a *grounding strategy*.

If a benchmark strategy was to have any last chance of rehabilitating an anthropocentric empiricism, therefore, it needed to be supplemented with a grounding strategy. Thus, Chapter 3 identified the Grounded Benchmark Criterion (GBC) as the best chance for any sort of benchmark approach to succeed. The GBC specified that we can observe via an instrument if and only if the apparatus is similar to human perception with respect to those features that make human perception (HP) reliable. However, as Chapter 3 showed, no sort of physical or causal similarity to human perception is a necessary condition for epistemic similarity to human perception. Instruments that bear no other resemblance to visual perception can share its reliability. Thus there is no unique epistemic privilege associated with being physically or causally similar to HP and both all forms of benchmark strategy and the anthropocentric version of empiricism must fail.

Having established the key role of reliability is grounding not just epistemically privileged forms of data production, but empiricism itself, Chapter 4 reviewed several existing accounts of reliability before going on to develop a novel account of reliability. This new account both provides a means to refine the weaker version of empiricism – showing why traditional empiricists were right about what they got right while providing the justification for eliminating what was wrong about the anthropocentric version – and allows us to assess the epistemic status of mathematized imaging technologies. The key idea developed in this chapter was that of *granularity*. Granularity is a characteristic of representations. We can refer to both the granularity of the world at which a particular question is directed and at the granularity of the representation (data) than an instrument generates. The granularity of a representation is the

smallest object or unit required to address the question of interest. The larger the spatial or temporal scale of the world which must be distinguished in order to answer a particular question, the larger the grain. The granularity of an instrument matches or is sufficient for a question if it is capable of providing evidence about the smallest objects needed to address that question. In many cases, the granularity of an instrument will be equal to its resolution. However, there are interesting and important cases where the granularity of an instrument is greater than its resolution (i.e. one can distinguish objects below the resolution of the instrument), so it is important to distinguish between the two. While it is not always possible to determine whether or not there is a granularity match in a given case, the latter part of Chapter 4 showed that there are a wide range of techniques available in order to establish the reliability of PET for particular applications.

Chapter 5 examined the significance of the fact that, though other data display formats are possible, the output of mathematized imaging technologies is usually images. While earlier chapters addressed epistemic features of the relationship between the object and its representation (i.e. the mode of production of the data), here I addressed the relationship between the representation and the viewer. In particular, Chapter 5 was concerned with the question of whether or not there is any epistemic privilege associated with certain data formats. Two possibilities were considered: 1) that images provide increased cognitive accessibility compared to other data formats, and 2) that moving images (videos) facilitate the identification of causal relationships. Cognitive accessibility was found to be an important feature of images. While it often results in a loss of effective granularity, presenting data in the form of an image does often increase the ease with which we can discriminate relevant features of the data. The identification of causal relationships, however, was not found to be affected by the data display format. If

videos were to provide special access to causal information, it must be that they allow us to perceive causation in a way that we cannot in static images or other data display formats. But the causal relationships that are even potentially visible with current imaging technologies are not the sorts of causes that we want to ascribe to the objects under study. We may see a cell move in a certain pattern and claim that it is chemotaxing, for instance. This is not a claim that is based only on our observation of the motion, but is based on background information about the type of cell it is and the presence of some unobserved chemoattractant in the medium. Thus, what we get from the data itself, however presented, is only spatiotemporal information, and without additional background information acquired in other ways we cannot identify specific causal relationships. However, it is often the case that the set of methods that are used in conjunction with video data collection (e.g visualizing specific proteins in a cell by creating fusions with green fluorescent protein and imaging living cells) are able to produce finer-grained data than do other methods where the data is represented as static images. As such, they may in practice (if not in theory) have a higher degree of granularity and so be able to more reliably answer certain sorts of questions.

This dissertation has shown that philosophical difficulties in understanding these technologies can not only be overcome, but than in the process we are led to a better understanding of the relationship between observation and evidence. At the end of this inquiry, we are left not only with the means to better assess the epistemic status of mathematized imaging technologies and to show why they can often produce very reliable evidence, but also with a more subtle version of empiricism that neither unduly privileges unaided human sense experience nor lacks the substance to distinguish between good and bad instruments or applications of instruments.

APPENDIX A

The Human Visual System

Light enters the eye by first passing through the cornea, which first begins to focus it. The light then passes through to the retina at the back of the eye. The retina converts light into neural signals. It consists of three layers composed of five types of cells – photoreceptors, bipolar cells, horizontal cells, amacrine cells, and ganglion cells – that collect light and extract basic information about color, form and motion. Incoming light travels through the other layers to reach the photoreceptor cells in the back. Photoreceptors are divided into two types – rods and cones. Rod cells are very sensitive to changes in contrast even at low light levels, but, as a result, are imprecise in detecting position (due to light scatter) and insensitive to color. They are primarily located in the periphery of the retina. Cones are high-precision cells that are specialized to detect red, green, or blue light. They are generally located in the center of the retina in an area of high spatial acuity called the fovea. Signals from the photoreceptor cells pass forward into the next layer of the cell containing horizontal, bipolar, and amacrine cells. These cells form small networks that are able to extract information about form and motion. That information continues on to the front of the retina where it is received by a layer of ganglion cells. The ganglion cells send out long, thin fibers that bundle together and go back down through the retina and out the back of the eye into the optic nerve. The spot where the optic nerve exits each eye has no light receptor cells and forms a blind spot in each eye.

The optic nerves within each eye meet in the front part of the head at a point called the optic chiasm. From there, all the fibers from the left half of each retina turn towards the right side of

the brain, and the fibers from the rights half of each retina head towards the left side of the brain. A small group of fibers in the optic nerve splits off and travels to the brainstem nuclei, which are groups of cells that govern reflex actions. Those fibers mediate automatic responses such as adjusting the size of the pupil, blinking, and coordinating the movement of the eyes. The majority of the fibers in the optic nerve, however, connect to a part of the occipital lobe called the primary visual cortex, or V1 (also known as the striate cortex).

On the way to V1, these fibers enter a part of the thalamus called the lateral geniculate nucleus (LGN), a layered structure with cells that respond to form motion and color. After fibers from the optic nerve enter the LGN, these streams of information are further separated before being sent on to V1. The connections from the eyes to the LGN and from the LGN to the V1 are topographically organized. This means that the mapping of each structure to the next is systematic: as you move along the retina from one point to another, the corresponding points in the LGN or V1 trace a continuous path. For example, the optic nerve fibers from a given small part of the retina all go to a particular small part of the LGN, and fibers from a given region of the LGN all go to a particular region of the primary visual cortex. In the retina, the successive stages are in apposition, so that the fibers can take a very direct route from one stage to the next. The cells in the LGN are obviously at a distance from the retina, as is V1 in a different place from the LGN. The style of connectivity nevertheless remains the same, with one region projecting to the next as though the successive areas were still superimposed. The optic-nerve fibers are gathered into a bundle as they leave the eye, and when they reach the LGN, they fan out and end in a topographically orderly way. Fibers leaving the LGN similarly fan out into a broad band that extends back through the interior of the brain and ends in an equally orderly way in the primary visual cortex. After several synapses, when fibers leave the primary visual cortex

and project to several other cortical regions, the topographic order is again preserved. Because convergence occurs at every stage, receptive fields tend to become larger: the farther along the path, the fuzzier this representation-by-mapping of the outside world becomes. An important, long-recognized piece of evidence that the pathway is topographically organized comes from clinical observation. If you damage a certain part of your primary visual cortex, you develop a local blindness, as though you had destroyed the corresponding part of your retina. The visual world is thus systematically mapped onto the LGN and cortex.

V1 is responsible for creating the basis of a three-dimensional map of visual space and for extracting features about the form and orientation of objects. Once basic processing has occurred in V1, the visual signal enters the secondary visual cortex, V2, which surrounds V1. V2 is primarily responsible for perceiving color and the relationships between form and color. V2 and higher cortical areas are generally referred to as extrastriate areas. Most of what we consider visual perception occurs in these extrastriate areas. They perform the two broad tasks of perceiving *what* forms are in the visual image and *where* objects are in space. The "what" tasks correspond to a number of connections in the temporal lobe (at the front of the brain, thus, the "what" stream is also known as the ventral stream), which contains areas that recognize objects and faces. The "where" tasks are performed through the dorsal stream into the parietal lobe, which has areas dedicated to perceiving motion and spatial relationships.

These "what" and "where" streams can be understood to each comprise a number of fairly general sub-modalities or functions. These different functions are listed in Table 7.1. There is an enormous amount of evidence from multiple areas of investigation that the visual system analyses the scene before it along these different dimensions in specialized modules that act in

parallel[195] beginning at the level of retinal cells and continuing through the various areas of the brain involved in vision.

Function[196]	**Basic requirement**
Photosensitivity	Photosensitive molecule
Form discrimination (spatial localization)	2-D array of photoreceptors plus focusing mechanism
Motion discrimination	2-D array of photoreceptors
Binocular vision and depth perception	Fusion of images
Color vision	Different photopigments

Table 6.1 Submodalities of vision.
Adapted from Shepherd (1988, 327).

These submodalities can be broken down into finer-grained functions which are associated with specific areas of the brain as is indicated in Figure 7.1. Even the most summary account of what is currently known about one of the submodalities in Figure 7.1 alone would fill dozens of pages and would involve descriptions of very particular processes and functions. This level of detail is not required for my project since the primary point to be

[195] Among researchers in early vision, this division of labor tends to be referred to as "parallelism", "multiplexing", or "partitioning", while "modularity" is the usual term used when referring to high level vision and visual cognition.
[196] I have left out discrimination of polarized light since this is not a submodality that human vision possesses, though other animals are able to make such discriminations and use it for orientation and navigation.

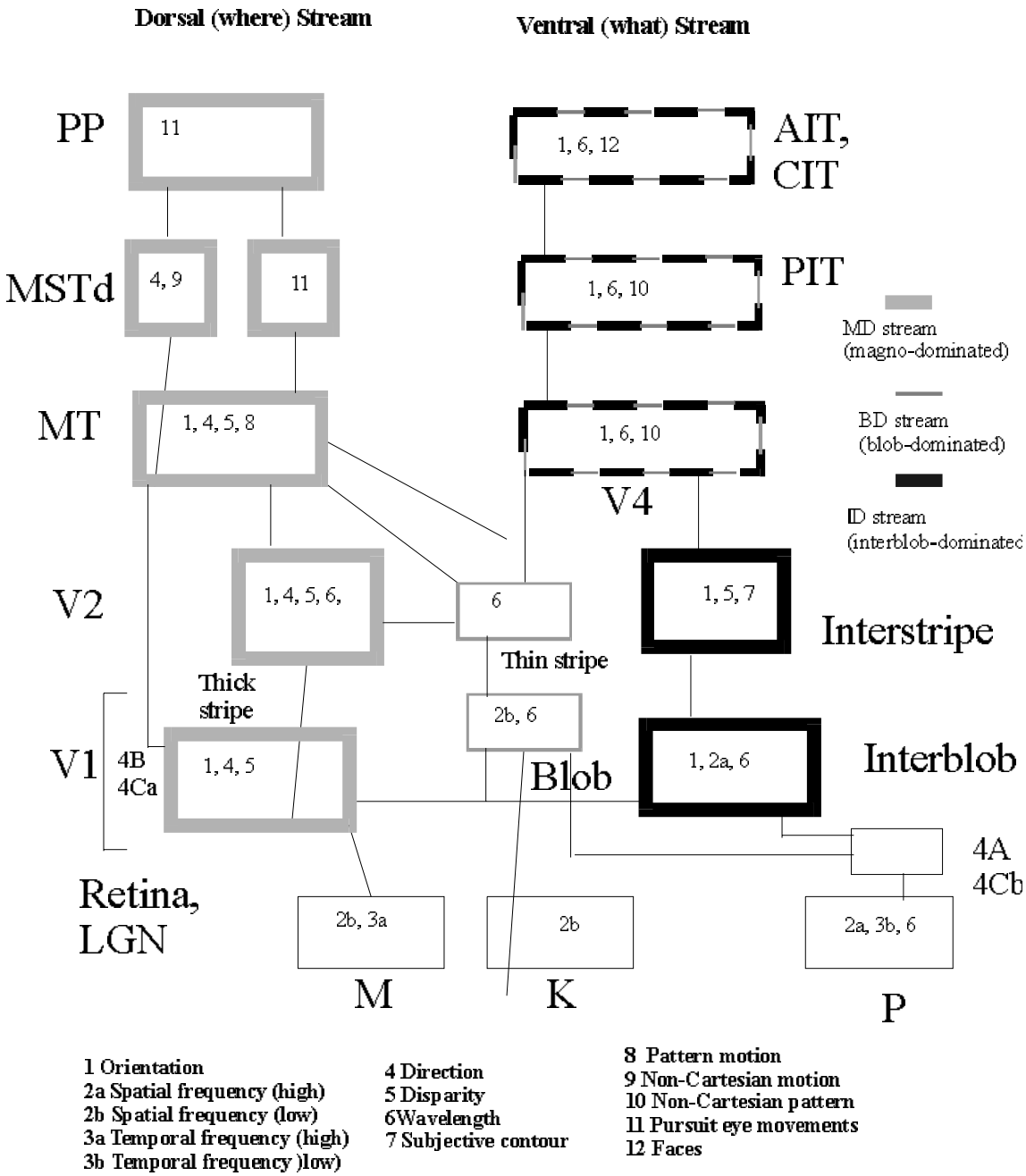


Figure 6.1 Hierarchical organization of the visual system in macaques.

Figure adapted from Van Essen and Gallant (1994).

made is that, although a great deal is known, there are still an enormous number of gaps in our knowledge.

The primary visual cortex is a part of the visual system which is relatively well understood from the point of view of neurophysiology. There exists a great deal of information about the different types of cells that are found in this area, the responses of these cells to different visual stimuli, and the neural connections and physical layout of these cells (see, for instance, Palmer 1999). Evidence from cognitive psychology – including the effects of lesions in various locations within the visual cortex as well as imaging studies that attempt to localize where different processes take place - have been very informative for identifying what different parts do and which functions can be dissociated from one another. What is less well understood, however, is *how* this area carries out these functions.[197] Providing a list of functions that appear to be carried out in this area (and that may or may not have analogues in specific imaging technologies) is possible, but this is clearly insufficient for assessing how or whether these functions contribute to the reliability of HP. (This question is not answered by the general veridicality of HP since many optical illusions are explained by the operation of the same functions that generally produce veridical perception. In other words, there is not a subset of functions that we can cordon off as contributing to the sometimes unreliability of HP - even under optimal conditions – and so leave the remainder as reliability-producing and therefore something that the GBC identifies as needing to be performed in a relevantly similar way in imaging technologies.) What we need is an account of how these functions are carried out in a

[197] Martha Farah, in summarizing the current state of knowledge, has the following to say: "…yet we still know relatively little about how this part of the brain [the primary visual cortex] subserves perception, in the sense of identifying functional perceptual mechanisms with the machinery described by neuroscientists. (…) The organization of the early visual cortices has been subject to intensive study in neuroscience, resulting in some hard-won and, in their own way, beautiful descriptions of visual anatomy and physiology. However, in many cases it has been impossible to assign any functional role to this organization, and when such attributions have been made they have been controversial." (2000, 20).

way that generates reliable information about the set of properties accessible to that form of imaging (whether HP or an instrument). Here, there are often multiple models that have been proposed to account for the empirical evidence. This multiplicity of mechanisms is then reproduced and expanded at the level of computational models.

BIBLIOGRAPHY

Abbott, A. (2003). A new atlas of the brain. *Nature, 424*, 249-250.

Abraham, T. (2003). From theory to data: Representing neurons in the 1940's. *Journal of the History of Biology*, 415-426.

Achinstein, P. (1985). *The Nature of Explanation*. New York: Oxford University Press.

Achinstein, P. (2000). Why philosophical accounts of evidence are (and ought to be) ignored by scientists. *Philosophy of Science, 67*, S180-S192.

Achinstein, P. (2001). *The Book of Evidence*. New York: Oxford University Press.

Alston, W. P. (1993). *The Reliability of Sense Perception*. Ithaca, N.Y.: Cornell University Press.

Ambrose, J. (1973). Computerized transverse axial tomography scanning (tomography): Part 2. Clinical application. *British Journal of Radiology, 46*, 1023-1047.

Amit, Y., Grenander, U., and Piccioni, M. (1991). Structural image restoration through deformable templates. *Journal of the American Statistics Association, 86*(414), 376-386.

Anscombe, G. E. M. (2004 [1971]). Causality and Determination, *Causation* (pp. 88-104). New York: Oxford University Press.

Arnheim, R. (1969). Visual Thinking. Berkeley: University of California Press.

Ashburner, J., and Friston, K.J. (1999). Nonlinear spatial normalization using basis functions. *Human Brain Mapping, 7*, 254-266.

Azzouni, J. (2004). Theory, observation and scientific realism. *British Journal for the Philosophy of Science, 55*, 371-392.

Baghaei, H., Wong. W.-H., Uribe, J., Li, H., Wang, Y., Liu, Y., Xing, T., Ramirez, R., Xie, S., and Kim, S. (2004). A comparison of four image reconstruction algorithms for 3-D PET imaging of MDAPET camera using phantom data. *IEEE Transactions on Nuclear Science, 51*(5), 2563-2569.

Bai, C., Kinahan, P.E., Brasse, D., Comtat, C., Townsend, D.W., Meltzer, C.C., Villemagne, V., Charron, M., and Defrise, M. (2003). An analytic study of the effects of attenuation on tumor detection in whole-body PET oncology imaging. *Journal of Nuclear Medicine, 44*, 1855-1861.

Baigrie, B. (Ed.). (1996). *Picturing knowledge: Historical and philosophical problems concerning the use of art in science*. Toronto: University of Toronto Press.

Bajcsy, R., and Kovacic, S. (1989). Multiresolution elastic matching. *Computer Vision, Graphics, and Information Processing, 46*, 1-21.

Barrington, S. F., and O'Doherty, M.J. (2003). Limitations of PET for imaging lymphoma. *European Journal of Medical Molecular Imaging, 30*(supplement), S117-S127.

Barsalou, L. W. (1992). *Cognitive Psychology: An Overview for Cognitive Scientists*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Beckman, M. (2003). Play-by-play imaging rewrites cells' rules. *Science, 300*, 76-77.

Beebe, J. (2004). The generality problem, statistical relevance and the tri-level hypothesis. *Nous, 38*, 177-195.

Bengel, F. M., Ziegler, S.I., and Avril, N. (1997). Whole body positron emission tomography in clinical oncology: comparison between attenuation corrected and uncorrected images. *European Journal of Nuclear Medicine, 24*, 1091-1098.

Berrill, N. J. (1984). The pearls of wisdom: An exposition. *Perspectives in Biology and Medicine, 28*(1), 1-16.

Bleckman, C., Jorg, D., Bohuslavizki, K.H., et al. (1999). Effect of attenuation correction on lesion detectability in FDG PET of breast cancer. *Journal of Nuclear Medicine, 40*, 2021-2024.

Bogen, J. (2001). Functional Imaging Evidence: Some Epistemic Hot Spots. In P. K. Machamer, Grush, R., and McLaughlin, P. (Ed.), *Theory and Method in the Neurosciences* (pp. 173-199). Pittsburgh: University of Pittsburgh Press.

Bogen, J. (2002). Epistemological custard pies from functional brain imaging. *Philosophy of Science, 69*(Supplement), S59-S71.

Bogen, J., and Woodward, J. (1988). Saving the Phenomena. *Philosophical Review, 97*, 303-357.

Bogen, J., and Woodward, J. (1992). Observations, theories and the evolution of the human spirit. *Philosophy of Science, 59*, 590-611.

Bookstein, F. L. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactionsof Pattern Analysis and Machine Intelligence, 11*, 567-585.

Brandom, B. (1995). Knowledge and the Social Articulation of the Space of Reasons. *Philosophy and Phenomenological Research, 55*, 895-908.

Breidbach, O. (2002). Representation of the Microcosm - The Claim for Objectivity in 19th Century Scientific Mirophotography. *Journal of the History of Biology, 35*, 221-250.

Brooks, R. A., and Di Chiro, G. (1976). Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging. *Physics in Medicine and Biology, 21,* 689-732.

Brown, R. S., Leung, J.Y., Kison, P.V., Zasadny, K.R., Flint, A., and Wahl, R.L. (1999). Glucose transporters and FDG uptake in untreated primary human non-small cell lung cancer. *Journal of Nuclear Medicine, 40,* 556-565.

Carnap, R. (1956). *Meaning and Necessity* (2nd ed.). Chicago: University of Chicago Press.

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago: University of Chicago Press.

Cartwright, L. (1995). *Screening the body: Tracing medicine's visual culture*. Minneapolis: University of Minnesota Press.

Christensen, G. E., Joshi, S.C., and Miller, M.I. (1997). Volumetric transformation of brain anatomy. *IEEE Transactions in Medical Imaging, 16*(6), 864-877.

Cohen, J., and Meskin, A. (2004). On the epistemic value of photographs. *The Journal of Aesthetics and Art Criticism, 62*(2), 197-210.

Collins, D. L., Evans, A.C., Holmes, C., and Peters, T.M. (1995). Automatic 3D segmentation of neuroanatomical structures from MRI. *Proceedings of Annual Conference on Information Processing in Medical Imaging,* 139-152.

Collins, D. L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., and Evans, A.C. (1998). Design and construction of a realistic digital brain phantom. *IEEE Transactions in Medical Imaging, 17,* 463-468.

Conee, E., and Feldman, R. (1998). The generality problem for reliabilism. *Philosophical Studies, 89*(1), 1-29.

Cox, D. D., and Savoy, R.L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage, 19,* 261-270.

Crick, F. (1994). *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Charles Scribner's Sons.

Crivello, F., Schorman, T., Tzouzio-Mazoyer, N., Roland, P.E., Zilles, K., and Mazoyer, B.M. (2002). Comparison of spatial normalization procedures and their impact on functional maps. *Human Brain Mapping, 16,* 228-250.

Cummins, R. (1996). *Representation, targets, and attitudes*. Cambridge, Massachusetts: MIT Press.

Cutler, S. A., and Ehrhardt, D.W. (2002). Polarized cytokinesis in vacuolate cells of Arabidopsis. *Proceedings of the National Academy of Sciences USA, 99*(5), 2812-2817.

Daisne, J.-F., Sibomana, M., Bol, A., Doumont, T., Lonneux, M., and Gregoire, V. (2003). Tr-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. *Radiotherapy and Oncology, 69*, 247-250.

Daston, L. and Galison, P. (1992). The Image of Objectivity. *Representations, 40*, 81-128.

Davatzikos, C., Vaillant, M., Resnick, S.M., Prince, J.L., Letovsky, S., and Bryan, R.N. (1996). A computerized approach for morphological analysis of the corpus callosum. *Journal of Computer Assisted Tomography, 20*, 88-97.

De Valois, K. K. (Ed.). (2000). *Seeing*. San Diego, CA: Academic Press.

DeRose, K. (1999). Contextualism: An explanation and defense. In J. Greco, and Sosa, E. (Ed.), *The Blackwell Guide to Epistemology* (pp. 187-205). New York: Blackwell.

Derrington, A. (2000). Seeing Motion. In K. K. De Valois (Ed.), *Seeing* (pp. 259- 309). San Diego, California: Academic Press.

Dowe, P. (2000). *Physical causation*. Cambridge: Cambridge University Press.

Dretske, F. (1969). *Seeing and Knowing*. Chicago: University of Chicago Press.

Ducasse, C. J. (1994 [1926]). On the Nature and Observability of the Causal Relation. In E. Sosa, and Tooley, M. (Ed.), *Causation* (pp. 125-136). New York: Oxford University Press.

Dumit, J. (2004). *Picturing Personhood: Brain Scans and Biomedical Identity*. Princeton: Princeton University Press.

Eddy, W. F., and Young, T.K. (2000). Optimizing the resampling of registered images. In I. N. Bankman (Ed.), *Handbook of Medical Imaging: Processing and Analysis* (pp. 603-612). San Diego: Academic Press.

Edelman, S. (1999). *Representation and Recognition in Vision*. Cambridge, MA: MIT Press.

Elkins, J. (1999). *Pictures of the body: Pain and metamorphosis*. Stanford: Stanford University Press.

Erasmus, J. J., Connolly, J.E., McAdams, H.P., and Roggli, V.L. (2000). Solitary pulmonary nodules. I. Morphologic evaluation for differentiation of benign and malignant lesions. *RadioGraphics, 20*, 43-58.

Evans, A. C., Collins, D.L., Neelin, P., MacDonald, D., Kamber, M., and Marrett, T.S. (1994). Three-dimensional correlative imaging: applications in human brain mapping. In R. W. Thatcher, Hallett, M., Zeffiro, T., John, E.R., and Huerta, M. (Ed.), *Functional neuroimaging: Technical foundations* (pp. 145-162). Orlando, FL: Academic Press.

Farah, M. J. (2000). *The Cognitive Neuroscience of Vision*. Malden, Massachusetts: Blackwell.

Farquhar, T. H., Llacer, J., Hoh, C.K., et al. (1999). ROC and localization ROC analyses of lesion detection in whole-body FDG PET: effects of acquisition mode, attenuation correction and reconstruction algorithm. *Journal of Nuclear Medicine, 40*, 2043-2052.

Feyerabend, P. (1978). *Against Method*. New York: Schocken.

Fox, P. T., Perlmutter, S., and Raichle, M.E. (1985). A stereotactic method of anatomical localization for positron emission tomography. *Journal of Computer Assisted Tomography, 8*, 141-153.

Franklin, A. (1986). *The neglect of experiment*. Cambridge: Cambridge University Press.

Gallagher, B. M., Fowler, J.S., Gutterson, N.I., MacGregor, R.R., Wan, C.N., and Wolf, A.P. (1978). Metabolic trapping as a principle of radiopharmaceutical design: some factors responsible for the biodistribution of 18-F 2-deoxy-2-fluoro-D-glucose. *Journal of Nuclear Medicine, 19*, 1154-1181.

Gallant, J. L. (2000). The Neural Representation of Shape. In K. K. De Valois (Ed.), *Seeing* (pp. 311- 333). San Diego, California: Academic Press.

Geisler, W. S., and Albrecht, D.G. (2000). Spatial Vision. In K. K. De Valois (Ed.), *Seeing* (pp. 79- 128). San Diego, California: Academic Press.

Genovese, C. (2002). Thresholding of statistical maps. *NeuroImage*(15), 870-878.

Genovese, C. R., Lazar, N.A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage, 15*, 870-878.

Gentner, D. (2000). Are scientific analogies metaphors? Atlantic Highlands: Humanities Press.

Giere, R. (1996). Visual models and scientific judgment. In B.S. Baigrie (Ed.), *Picturing Knowledge: Historical and Philosophical Problems Concerning the Use of Art in Science* (pp. 269-302). Toronto: University of Toronto Press.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Ginet, C. (1985). Contra reliabilism. *Monist, 68*, 175-187.

Glenn, J. A., and Littler, G.H. (1984). *A dictionary of mathematics*. Towata, NY: Barnes and Noble.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis, 44*, 49-71.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*, S342-S353.

Glymour, C. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Glymour, C. (2003). Instrumental probability. In H. E. Kyburg, and Thalos, M. (Ed.), *Probability is the very guide of life: The philosophical uses of chance* (pp. 235-252). Chicago: Open Court.

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and knowledge: New studies in epistemology* (pp. 1-23). Dordrecht: Riedel.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goldman, A. (1994). Naturalistic epistemology and reliabilism. *Midwest Studies in Philosophy, 19*, 301-320.

Goldman, A. (1999). *Knowledge in a social world*. Oxford: Clarendon Press.

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Indiana: Hackett Publishing Company.

Gordon, R., Herman, G.T., and Johnson, S.A. (1975). Image reconstruction from projections. *Scientific American, 233*, 56-68.

Gould, M. K., Maclean, C.C., Kuschner, W.G., Rydzak, C.E., and Owens, D.K. (2001). Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: A meta-analysis. *JAMA, 285*(7), 914-924.

Greene, M. T. (2005). Seeing clearly is not necessarily believing. *Nature, 435*, 143.

Griesemer, J. R. (1991). Material models in biology. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science, 1990*, 79-93.

Grush, R. (2001). The semantic challenge to computational neuroscience. In P. K. Machamer, McLaughlin, P., and Grush, R. (Ed.), *Theory and Method in the Neurosciences*. Pittsburgh: University of Pittsburgh Press.

Haack, S. (1993). *Evidence and Inquiry: Towards Reconstruction in Epistemology*. Oxford: Blackwell.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Hammer, E. (1995). *Logic and visual information*. Stanford, CA: CSLI Publications.

Hanson, N. R. (1965). *Patterns of Discovery*. Cambridge: Cambridge University Press.

Harrell, M. (2000). *The reliability of methods for chaotic data analysis.* Doctoral dissertation. University of California, San Diego.

Haynes, J.-D., and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8*, 1-6.

Hearst, J. E. (1990). Microscopy:'Seeing is Believing'. *Nature, 347*(6290), 230.

Hempel, C. (1965). *Aspects of scientific explanation*. New York: Free Press.

Herschman, H. R., MacLaren, D.C., Iyer, M. et al. (2000). Seeing Is Believing: Non-invasive, Quantitative and Repetitive Imaging of Reporter Gene Expression in Living Animals Using Positron Emission Tomography. *Journal of Neuroscience Research, 59*(6), 699-705.

Hillyard, S. A., Vogel, E.K., and Luck, S.J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 393*, 1257-1270.

Hoffmann, E. J., Cutler, P.D., Digby, W.M., et al. (1990). 3-D phantom to simulate cerebral brain flow and metabolic images for PET. *IEEE Transactions in Nuclear Science, 37*, 616-620.

Hofstadter, D. (1979). *Godel, Escher, Bach*. New York: Basic Books.

Hsu, C.-H. (2002). A study of lesion contrast recovery for iterative PET image reconstructions vesus filtered backprojection using an anthropocentric thoracic phantom. *Computerized Medical Imaging and Graphics, 26*, 119-127.

Hume, D. (1980 [1740]). *A treatise of human nature* (2nd, revised ed.). Oxford: Clarendon Press.

Humphreys, P. (2004). *Extending ourselves: Compuational science, empiricism, and scientific method*. New York: Oxford University Press.

Hurtley, S. M., and Helmuth, L. (2003). The future looks bright ... *Science, 300*, 75.

Imdahl, A., Jenkner, S., Brink, I., Nitzsche, E., Stoelben, E., Moser, E., and Hasse, J. (2001). Validation of FDG positron emission tomography for differentiation of unknown pulmonary lesions. *European Journal of Cardiothoracic Surgery, 20*, 324-329.

Jones, C. A., and Galison, P (Ed.). (1988). *Picturing science, producing art*. New York: Routledge.

Judson, H. F. (1981). *The search for solutions*. New York: Holt, Rinehart and Winston.

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8*, 679-685.

Kandel, E. R., Schwartz, J.H., and Jessell, T. (Ed.). (2000). *Principles of Neural Science* (4th ed.). New York: McGraw-Hill.

Keeley, B. L. (1999). Fixing content and function in neurobiological systems: The neuroethology of electroreception. *Biology and Philosophy, 14*, 395-430.

Keeley, B. L. (2002). Making sense of the senses. *Journal of Philosophy, 99*, 5-28.

Keller, E. F. (2002). *Making Sense of Life: Explaining Biological Development With Models, Metaphors, and Machines*. Cambridge, Massachusetts: Harvard University Press.

Kelly, K. T. (1996). *The Logic of Reliable Inquiry*. New York: Oxford University Press.

Kelly, K. T. (2004). Justification as truth-finding efficiency: How Ockham's razor works. *Minds and Machine*, 1-21.

Kelly, K. T., Schulte, O., and Juhl, C. (1997). Learning theory and the philosophy of science. *Philosophy of Science, 64*, 245-268.

Kersten, D., Mamassian, D., and Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304.

Kevles, B. H. (1996). *Medical imaging in the twentieth century*. Brunswick, NJ: Rutgers University Press.

Kim, B., Boes, J.L., Frey, K.A., and Meyer, C.R. (1997). Mutual information for automated unwarping of rat`brain autoradiographs. *NeuroImage, 5*(1), 31-40.

Kim, D. S., Duong, T.Q., and Kim, S.-G. (2000). High-resolution mapping of iso-orientation columns by fMRI. *Nature Neuroscience, 3*(2), 164-169.

Kim, J. (1995). Causation. In R. Audi (Ed.), *The Cambridge Dictionary of Philosophy* (pp. 110-112). Cambridge: Cambridge University Press.

Kitcher, P., and Varzi, A. (2002). Some pictures are worth 2 χ_o sentences. *Philosophy, 75*, 277-381.

Knoll, G. G. (1989). *Radiation detection and measurement*. New York: John Wiley and Sons.

Kosso, P. (1992). Observation of the past. *History and theory, 31*(1), 21-36.

Koyama-Honda, I., Ritchie, K., Fujiwara, T., Iino, R., Murakoshi, H., Kasai, R.S., and Kusumi, A. (2005). Fluorescence imaging for monitoring the colocalization of two single molecules in living cells. *Biophysical Journal, 88*(3), 2126-2136.

Kuhn, T. (1970). *The Structure of Scientific Revolutions* (Second, Enlarged ed.). Chicago: University of Chicago Press.

Ladyman, J. (2002). *Understanding Philosophy of Science*. London: Routledge.

Lauf, U., Giepmans, B.N.G., Lopez, P., Braconnot, S., Chen, S.-C., and Falk, M.M. (2002). Dynamic traficking and delivery of connexons to the plasma membrane and accretion to gap junctions in living cells. *Proceedings of the National Academy of Sciences USA, 99*(16), 10446-10451.

Lee, T. S. (2003). Computation in the early visual cortex. *Journal of Physiology - Paris, 97*, 121-139.

Lewitt, R. M., and Matej, S. (2003). Overview of methods for image reconstruction from projections in emission computed tomography. *Proceedings of the IEEE, 91*(10), 1588-1611.

Liou, M., Lee, J.-D., Cheng, P.E., Huang, C.-C., and Tsai, C.-H. (2003). Bridging functional MR images and scientific inference: Reproducibility maps. *Journal of Cognitive Neuroscience, 15*, 935-945.

Lippincott-Schwartz, and Patterson, G.H. (2003). Development and use of fluorescent protein markers in living cells. *Science, 300*, 87-90.

Locke, J. (1979 [1690]). *An essay concerning human understanding*. Oxford: Oxford University Press.

Lommerse, P. H. M., Snaar-Jagalska, B.E., Spaink, H.P., and Schmidt, T. (2005). Single-molecule diffusion measurements of H-Ras at the plasma membrane of live cells reveal microdomain localization upon activation. *Journal of Cell Science, 119*(9), 1799-1809.

Lukic, A. S., Wernick, M.N., and Strother, S.C. (2002). An evaluation of methods for detecting brain activations from functional neuroimages. *Artificial Intelligence in Medicine, 25*, 69-88.

Lynch, M., and Woolgar, S. (Ed.). (1990). *Representation in Scientific Practice*. Cambridge, MA: MIT Press.

Machamer, P. K. (1970). Observation. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science, 1970*, 187-201.

Machamer, P. K. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science, 18*(1), 27-39.

Machamer, P. K., Darden, L., and Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Matthies, A., Hickeson, M., Cuchiara, A., and Alavi, A. (2002). Dual time point 18F-FDG PET for the evaluation of pulmonary nodules. *Journal of Nuclear Medicine, 43*(7), 871-875.

Maxwell, G. (1962). The ontological status of theoretical entities. In H. Feigl, and Maxwell, G. (Ed.), *Minnesota Studies in the Philosophy of Science* (Vol. 3, pp. 3-15). Minneapolis, MN: University of Minnessota Press.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. (2000). Experimental practice and an error statistical account of evidence. *Philosophy of Science, 67*, S193-S207.

McGinn, C. (1999). *Knowledge and Reality*. Oxford: Clarendon Press.

McIntosh, A. R., and Lobaugh, N.J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. *NeuroImage, 23*, S250-S263.

Menuge, A. (1995). The scope of observation. *Philosophical Quarterly, 45*, 60-69.

Michotte, A. (1963[1946]). *The perception of causality* (T. Miles, and Miles, E., Trans.). New York: Basic Books.

Miller, R. G. J. (1981). *Simultaneous Statistical Inference* (2nd ed.). New York: Springer-Verlag.

Mitchell, S. D. (1997). Pragmatic Laws. *Philosophy of Science, 64*(Proceedings), S468-S479.

Mitchell, S. D. (2000). Dimensions of Scientific Law. *Philosophy of Science, 67*, 242-265.

Mitchell, S. D. (2002). Integrative Pluralism. *Biology and Philosophy, 17*, 55-70.

Mohler, W. A. (1999). Visual Reality: Using Computer Reconstruction and Animation to Magnify the Microscopists's Perception. *Molecular Biology of the Cell, 10*, 3061-3065.

Monga, O., and Benayoun, S. (1995). Using partial derivatives of 3D surfaces to extract typical surface features. *Computer Vision and Image Understanding, 61*(2), 171-189.

Monteith, G. R. (2000). Seeing Is Believing: Recent Trends in the Measurement of Ca2 in Subcellular Domains and Intracellular Organelles. *Immunology and Cell Biology, 78*(4), 403-407.

Murakoshi, H., Iino, R., Kobayashi, T., Fujiwara, T., Ohshima, C., Yoshimura, A., and Kusami, A. (2004). Single-molecule imaging analysis of Ras activation in living cells. *Proceedings of the National Academy of Sciences USA, 101*, 7317-7322.

Murphy, B. (1996). *Color scales: dialing a defect*. Retrieved January 16, 2005, from the World Wide Web: www.nucmed.buffalo.edu/nrlgy1.htm

Norton, J. D. (2003). *Causation as folk science*. Retrieved, 2003, from the World Wide Web: www.philosophersimprint.org/003004

Norton, J. D. (2004). Why thought experiments do not transcend empiricism. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 44-66). Malden, MA: Blackwell.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Belknap Press.

Orr-Weaver, T. L. (1995). Meiosis in Drosophila: Seeing Is Believing. *Proceedings of the National Academy of Sciences USA, 92*(23), 10443-10449.

Osman, M. M., Cohade, C., Nadamoto, Y., and Wahl, R. (2003). Clinically significant inaccurate localization of lesions with PET-CT: Frequency in 300 patients. *Journal of Nuclear Medicine, 44*(2), 240-243.

Ottino, J. A. (2003). Is a picture worth 1,000 words? *Nature, 421*, 474-476.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, Massachusetts: MIT Press.

Parker, A. J., and Newsome, W.T. (1998). Sense and the Signal Neuron: Probing the Physiology of Perception. *Annual Review of Neuroscience, 21*, 227-277.

Patz, E. F. J., Vowe, V.J., Hoffman, J.M., Paine, S.S., Burrowes, P., Coleman, R.E., and Goodman, P.C. (1993). Focal pulmonary abnormalities: Evaluation with F-18 fluorodeoxyglucose PET scanning. *Radiology, 188*(2), 487-490.

Pearson, H. (2005). CSI: cell biology. *Nature, 434*, 952-953.

Perini, L. (2002). *Visual representations and scientific knowledge.* Doctoral dissertation. University of California, San Diego.

Perini, L. (2005). The truth in pictures. *Philosophy of Science, 72,* 262-285.

Phelps, M. E., Hoffmann, E.J., Mulliani, et al. (1975). Application of annihilation coincidence detection to transaxial reconstruction tomography. *Journal of Nuclear Medicine, 16,* 210-224.

Poeppel, D., and Embick, D. (forthcoming). Defining the relation between linguistics and neuroscience. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones*. Hillsdale, NJ: Lawrence Erlbaum.

Price, H. H. (1950). *Perception* (2nd, revised ed.). Westport, Connecticut: Greenwood Press.

Pryor, J. (2001). Highlights of recent epistemology. *British Journal for the Philosophy of Science, 52*, 95-124.

Putnam, H. (1975). The meaning of 'meaning', *Mind, Language and Reality* (pp. 215-271). Cambridge: Cambridge University Press.

Putnam, H. (1975). *Mind, Language, and Reality: Philosophical Papers* (Vol. 2). Cambridge: Cambridge University Press.

Pylyshyn, Z. (2003). Return of the mental image: are there really pictures in the brain? *Trends in Cognitive Sciences, 7*(3), 113-118.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60,* 20-43.

Rasmussen, N. (1997). *Picture control: The electron microscope and the transformation of biology in America, 1940-1960*. Stanford: Stanford University Press.

Reynolds, J. H., and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron, 24*, 19-29.

Rieder, C. L., and Khodjakov, A. (2003). Mitosis through the microscope: Advances in seeing inside live dividing cells. *Science, 300*, 91-96.

Ritchie, K., Shan, X.-Y., Kondo, J., Iwasawa, K., Fujiwara, T., and Kusumi, A. (2005). Detection of non-Brownian diffusion in the cell membrane in single molecule tracking. *Biophysical Journal, 88*(3), 2266-2277.

Rockland, K. S., and Pandya, D.N. (1979). Laminar origins and terminations of cortical connections of the occipetal lobe in the rhesus monkey. *Brain Research, 179*, 3-20.

Rohren, E. M., Turkington, E.M., and Colemn, R.E. (2004). Clinical Applications of PET in Oncology. *Radiology, 231*, 305-332.

Roush, S. (2005). *Tracking truth: Knowledge, evidence, and science*. New York: Oxford University Press.

Rudge, D. W. (1998). A Bayesian analysis of strategies in evolutionary biology. *Perspectives on Science, 6*(4), 341-360.

Rudge, D. W. (1999). Taking the peppered moth with a graimn of salt. *Biology and Philosophy, 14*(1), 9-37.

Ruse, M. (1991). Are pictures really necessary? The case of Sewall Wright's "Adaptive Landscapes". *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science, 1990*, 63-77.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Sanchez-Crespo, A., Andreo, P., and Larsson, S.A. (2004). Positron flight in human tissues and its influence on PET image spatial resolution. *European Journal of Medical Molecular Imaging, 31*, 44-51.

Sargent, P. (1996). On the use of visualization in the practice of science. *Philosophy of Science, 63*, S230-S238.

Scannell, J. W., Blakemore, C., and Young, M.P. (1995). Analysis of connectivity in the cat cerebral cortex. *Journal of Neuroscience, 15*, 1463-1483.

Schlottmann, A. (2001). Perception versus knowledge of cause and effect in children: When seeing is believing. *Current Directions in Psychological Science, 10*, 111-115.

Schlottmann, A., Allen, D., Linderoth, C., and Hesketh, S. (2002). Perceptual causality in children. *Child Development, 73*, 1656-1677.

Schoenahl, F., Montandon, M.L., Slosman, D.O., and Zaidi, H. (2003). Assessment of the performance of SPM analysis in PET neuroactivation studies: A Monte Carlo investigation.

Scholl, B. J., and Nakayama, K. (2002). Causal capture: Contextual effects on the perception of collision events. *Psychological Science, 13*(6), 493-498.

Scholl, B. J., and Nakayama, K. (2004). Illusory causal crescents: Misperceived spatial relations due to perceived causality. *Perception, 33*, 455-469.

Seisenberger, G., Ried, M.U., Endress, T., Buning, H., Hallek, M., and Brauchle, C. (2001). Real-time single-molecule imaging of the infection pathway of an adeno-associated virus. *Science, 294*, 1929-1932.

Shapere, D. (1982). The concept of observation in science and philosophy. *Philosophy of Science, 49*, 485-525.

Shapley, R. (2000). Receptive Fields of Visual Neurons. In K. K. De Valois (Ed.), *Seeing* (pp. 55- 78). San Diego, California: Academic Press.

Shepard, G. M. (1988). *Neurobiology* (Second ed.). Oxford: Oxford University Press.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences, 8*(5), 223-230.

Spirtes, P. G., C., and Scheines, R. (2000). *Causation, Prediction, and Search* (2nd ed.). Cambridge, MA: MIT Press.

Stafford, B. M. (1991). *Body criticism:Imaging the unseen in Enlightenment art and medicine*. Cambridge, MA: MIT Press.

Stafford, B. M. (1994). *Artful science: Enlightenment entertainment and the eclipse of visual education*. Cambridge, MA: MIT Press.

Stafford, B. M. (1996). *Good looking: Essays on the virtues of images*. Cambridge, MA: MIT Press.

Steel, D. (2005). The facts of the matter: A discussion of Norton's material theory of induction. *Philosophy of Science, 72*.

Stephens, D. J., and Allan, V.J. (2003). Light microscopy techniques for live cell imaging. *Science, 300*, 82-86.

Stuffelbeam, R. S., and Bechtel, W. (1997). PET: Exploring the myth and the method. *Philosophy of Science, 64*, S95-S106.

Subsol, G. (1999). Crest lines for curve based warping. In A. W. Toga (Ed.), *Brain warping*. San Diego: Academic Press.

Supplee, C. and Bradford, M. (2004). 2004 Visualization Challenge. *Science, 305*, 1903.

Szeliski, R., and Lavallee S. (1993). Matching 3D anatomical surfaces with non-rigid deformations using octree-splines. *SPIE, 2031*.

Taylor, P. J. (1991). Mapping ecologists' ecologies of knowledge. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science, 1990*, 95-109.

Thibos, L. N. (2000). Formation and Sampling of the Retinal Image. In K. K. De Valois (Ed.), *Seeing* (pp. 1- 54). San Diego, California: Academic Press.

Thines, G., Costall, A., and Butterworth, G. (Ed.). (1991). *Michotte's Experimental Phenomenology of Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, P. M., and Toga, A.W. (1996). A surface-based technique for warping three-dimensional images of the brain. *IEEE Transactions in Medical Imaging, 15*(4), 402-417.

Toga, A. W. (1997). A deformable high resolution anatomic reference for PET activation studies. In B. Gulyas (Ed.).

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology in Human Perception and Performance, 8*, 194-214.

Treisman, A. (1998). Feature binding, attention, and object perception. *Transactions of the Royal Society of London: Group B Biological Sciences, 353*, 1295-1306.

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. *Neuron, 24*, 105-110.

Treisman, A., and Kanwisher, N.K. (1998). Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology, 8*, 218-226.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, Connecticut: Graphics Press.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, Connecticut: Graphics Press.

Turnbull, D. (1989). *Maps are territories*. Chicago: University of Chicago Press.

Twardy, C. R., and Bingham, G.P. (2002). Causation, causal perception, and conservation laws. *Perception & Psychophysics, 64*(6), 956-968.

Van Essen, D. C., and Gallant, J.L. (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron, 13*, 1-10.

van Fraassen, B. (1980). *The Scientific Image*. New York: Oxford University Press.

Van Orden, G. C., and Paap, K.R. (1997). Functional neuroimages fail to discover pieces of mind in the parts of the brain. *Philosophy of Science, 64*, S85-S94.

Vanduffel, W., Tootell, R.B., Schoups, A.A., and Orban, G.A. (2002). The organization of orientation selectivity throughout macaque visual cortex. *Cerebral Cortex, 12*, 647-662.

Vezoli, J., Falchier, A., Jouve, B., Knoblauch, K., Young, M., and Kennedy, H. (2004). Quantitative analysis of connectivity in the visual cotrex: Extracting function from structure. *Neuroscientist, 10*(5), 476-482.

Wahl, R. L. (1999). To AC or not to AC: that is the question. *Journal of Nuclear Medicine, 40*, 2025-2028.

Wahl, R. L., Neuhoff, A., Kison. P., and Zasadny, K.R. (1997). ROC and localization ROC analyses of lesion detection in whole-body FDG PET: effects of acquisition mode, attenuation correction, and reconstruction algorithm. *Journal of Nuclear Medicine, 39*(44P).

Wang, G. J., Volkow, N.D., Pappas, N.R., Wong, C.T., Nutsull, N., and Fowler, J.S. (2001). Brain dopamine and obesity. *The Lancet, 357*(9253), 354-357.

Warburg, O. (1930). *The metabolism of tumors*. London: Constabee.

Weijer, C. (2003). Visualizing signals moving in cells. *Science, 300*, 96-100.

White, P. A., and Milne, A. (1999). Impressions of enforced disintegration and bursting in the visual perception of collision events. *Journal of Experimental Psychology, 110*, 573-602.

White, P. A., and Milne, A. (2003). Visual impressions of penetration in the perception of objects in motion. *Visual Cognition, 10*(5), 605-619.

White, P. A., & Milne, A. (2003). Visual impressions of penetration in the perception of objects in motion. *Visual Cognition, 10*(5), 605-619.

Wimsatt, W. C. (1991). Taming the dimensions - Visualizations in science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science, 1990*, 111-135.

Wolfe, J. M., and Bennett, S. (1996). Preattentive object files: shapeless bundles of basic features. *Vision Research, 37,* 25-44.

Wolfe, J. M., and Cave, K.R. (1999). The psychophysical evidence for a binding problem in human vision. *Neuron, 24*, 11-17.

Wolfe, J. M., and Horowitz, T.S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neurosciience, 5*, 1-7.

Wolfe, J. M., Cave, K.R., and Franzel, S.L. (1989). Guided Search: an alternative to the Feature Integration model for visual search. *Journal of Experimental Psychology on Human Perception and Performance, 15*, 419-433.

Wolfe, J. M., Horowitz, T.S., Kenner, N.M. (2005). Rare items often missed in visual searches. *Nature, 435*, 439.

Woodward, J. (2000). Data, phenomena, and reliability. *Philosophy of Science, 67*, S163-S179.

Ziman, J. (1978). *Reliable Knowledge*. Cambridge: Cambridge University Press.