Facta Philosophica 5, 2003: 335–353
© Peter Lang, Switzerland

Two Realms of Mental Life: The Non-overlap of Belief Ascription and the Scientific Study of Mind and Behavior

Nick Chater & Martin J. Pickering

There are two, very different, ways of studying human nature. One approach is what we shall call *scientific* explanation. Here, the project is to understand the structure and function of the human brain, and related systems, in just the way we might attempt to understanding the brains of animals; or indeed, the physiology of the heart, the respiration of a tree, the behaviour of a tornado, or the operation of a computer. The fact that the student and the material studied happen to be of the same kind (i. e., people are studying their own nature) is incidental. A human is viewed as just one more, albeit rather complex, mechanical system. This pattern of explanation is standard in the brain and cognitive sciences. The other approach to understanding human nature treats people *as people*. It views people in terms of their beliefs, attitudes, desires and culture (Fodor, 1987). Here, the project is to attempt to enter the minds of others, to see the world from their point of view. And the fact that the student and the material studied happen to be of the same kind is of central importance–the imaginative leap required to understand others is proportional to their similarity to oneself; and the project of entering the minds of newborn babies or bats is challenging or perhaps even incoherent (Nagel, 1974); and one clearly cannot attempt to enter the mind of a heart or a tornado, because hearts and tornados don't have minds. Let us call this personal-level explanation–where people's thoughts and behavior are explained in terms of their propositional attitudes, and related concepts. Personal-level explanation is involved in our everyday descriptions of each other's thought and behavior; and it is the standard mode of explanation in literature and the humanities.

In principle, the two types of explanation appear to apply to the same phenomena. Suppose that the room becomes hot, and I open a window. There is a mechanistic story that underpins this event, which could be described, for example, at a neural level. A chain of extraordinarily complex sensory, neural and muscular events leads from the increase in ambient temperature, through a tangle of complex nervous activity, to a series of muscular contractions, that push the window open. From the standpoint of personal-level explanation, things are viewed differently: I notice (and hence come to believe) that it is hot; I wish (i. e., have the desire) to be less hot; I know various facts about windows, drafts, and air temperature (exactly how 'deep' this knowledge goes is not at issue here—I may just know that opening windows cools one down, with no background theory at all, of course); I am able to reason that opening the window provides a straightforward way to bring about my desire; so I decide to open it and act on this decision. So, it appears, the very same episode, can be explained in two very different ways.

How do these personal-level and scientific explanations relate to each other? One viewpoint is that these views are *competitive*: that they cannot both be right. There are, predictably, two versions of this view. One version is that personal-level explanation is correct; and that this fundamentally undermines the project of attempting to provide a scientific explanation of human nature (e. g., Shotter, 1975). The other version is that scientific explanation is correct, and will gradually drive out 'unscientific' personal-level explanations (e.g., Churchland, 1986).

The second viewpoint is that these explanations are *complementary*, rather than competitive. Complementary levels of explanation of the same phenomena have a respectable place in science. To take a standard example, the rate at which a gas becomes hotter when it is compressed can be explained at a macroscopic level by Boyle's Law, or at a microscopic level, by the mechanical interactions of many tiny particles. Of course, merely raising the possibility that two types of explanation are complementary is not enough—what is required is some account of how they can be compatible. In the case of gases, this was made possible by the development of statistical mechanics, which showed that the macro-level behaviour described by Boyle's Law emerges automatically from statistical aggregation of micro-level particles behaving according to Newton's laws. So how might the scientific and personal level explanations be reconciled? How can a world of causally interacting sensors and nerves support an explanation in terms of beliefs, desires, and actions? The putative analog of statistical mechanics in bridging between levels of explanation is the computational theory of mind (Fodor, 1974). The idea is that propositional attitudes are relations to internal mental representations. The *content* of these mental representations corresponds to the meaning of the *that*-clause in the propositional attitude; and the formal, computational properties of the representation (i. e., its causal powers) systematically 'track' this content. This 'tracking' ensures that the causal chains of representations, which are determined by the representation's formal properties, correspond to coherent propositional attitude explanations. So, to take a parallel with a simpler and better-understood case, in a computer program for proving logical theorems, the causal chains of representations generated by the program can be systematically interpreted as valid deductions. This approach to reconciling scientific and personal-level explanation has been widely influential in philosophy and the foundations of cognitive science (e. g., Fodor, 1987; Fodor & Pylyshyn, 1988). Moreover, it is the starting point for traditional symbolic approaches to artificial intelligence. Here, the goal is to elicit beliefs, desires and other propositional attitudes from people, and to attempt to codify these in a formal representational language, over which some type of theorem-proving mechanisms operate (e. g., Newell & Simon, 1972). According to this viewpoint, then, personal-level explanation is not rejected—but instead is at centre stage in developing scientific, computational models of thought and behavior.

There has been considerable debate concerning whether personal level and scientific explanation of the mind can be viewed as complementary in this way, or whether the two styles of explanation are better viewed as standing in competition (e. g., Churchland, 1986; Dreyfus & Dreyfus, 1986; Fodor, 1987; Stich, 1983).

Our thesis in this paper is that neither point of view, and, equally, the debate between them, is relevant to practical research in the brain and cognitive sciences, because, in practice, battle is never joined: there are no aspects of thought and behavior that can simultaneously be explained in both ways, and hence the question of whether such explanations are complementary or competitive never arises. There are, in short, two realms of cognitive phenomena—on the one side, phenomena that succumb to personal-level explanation; other the other, phenomena that succumb to scientific explanation. And these realms do not overlap.

We argue for our thesis in two ways, corresponding to the two main parts of this paper: first, we provide illustrative examples; and second, we provide a rationale for the existence of the divide.

Illustrating the divide

In this section, we aim to show that apparent overlap in scope between scientific and personal-level explanation is illusory. We contrast personal-level explanation with three kinds of scientific explanation from cognitive science, the brain sciences and rational choice in the social sciences.

Personal explanation and cognitive science

Personal-level explanation is concerned with the *content* of what people believe or desire. And content appears to have a huge influence on the operation of the mental processes. This is one way in which our putative divide appears to be crossed. Moreover the influence appears to go in the opposite direction: mental mechanisms patently influence personal-level explanation. Function, or misfunction, of perception, memory and reasoning, seem, inevitably, to influence our beliefs and desires—thus factors amenable to scientific explanation affect personal-level explanation.

We are arguing that there are two distinct realms of mental life—those that can be addressed by folk psychology and those that can be addressed by the neuro- and cognitive sciences. But this does not imply that these realms can be understood without reference to one another—that the scientist should ignore folk psychological explanation; or that folk psychology should be uninfluenced by the cognitive and neurosciences. This is because much human thought and behaviour is influenced by phenomena from both realms. We shall see below that intricate interactions between scientific and folk psychological explanation arise in many areas of research in cognition; and this interplay is perhaps one reason why the divide between the two approaches is not always apparent. Yet we claim that explanations involving both folk psychological and scientific components never involve providing different levels of analysis of the same phenomenon—rather the folk psychological and scientific components of explanation apply to non-overlapping, though potentially interacting, phenomena.

We begin with cognitive science, focussing on cases that appear to provide clear cases of phenomena which are simultaneously addressed by both scientific and personal-level explanation. We consider cases from study of memory and language processing in turn.

Memory

A ubiquitous finding in the cognitive psychology of memory is that the *content* of the materials being stored and retrieved has marked implications on memory performance. For example, the comprehensibility of a text dramatically affects recall (Bartlett, 1932). This is neatly demonstrated by presenting the same passage with or without a helpful explanatory title (Bransford & Johnson, 1972). Passages are chosen so that readers in the no-title condition are entirely baffled, while the passage is entirely cogent when the explanatory title is given. Memory for the passage is substantially impaired in the no-title condition.

Now what explains whether the passage is, or is not, comprehensible? The degree to which a passage is understood is substantially determined by the possession of, and ability to access, relevant background knowledge—and knowledge, a close relative of belief, is the province of personal-level explanation. Thus, with appropriate background beliefs (given by general background knowledge, and cued by the title), the reader can assimilate the sentences successfully. But where background knowledge is not appropriately cued (or is absent—as in the case of, say, reading about a ritual from an unfamiliar culture, or a theory an unfamiliar area of science), then memory is poor.

So it may appear that content, and thus personal-level explanation, occupies the same territory as scientific theories of memory. But there is no overlap—because there is no scientific psychological explanation of what makes the content of any specific passage comprehensible or not. Scientific explanation begins at the point where folk psychological explanation ceases: it takes as given that the content of some passages are more difficult to comprehend than others, and traces the implications of this for other aspects of cognition. For example, a scientific account might conjecture that a passage that is poorly understood does not lay down such a rich or cohesive representation; and there might be a computational theory concerning the mechanisms of memory storage, which explain why less cohesive memory traces are more difficult to recover, leading to poorer memory performance. But, crucially, scientific explanation does not trespass on the personal-level territory, of explaining *why* one passage is comprehensible, and another is not.

The same point arises for a host of psychological phenomena in memory. For example, suppose that a person is given a list of words to remember, one of which is in some way surprising or incongruous. Then, typically, that word will be remembered especially well—this is the von Restorff effect (von Restorff, 1933). But what is it about a word that makes it seems surprising, in a particular context? This appears to be the territory of personal-level explanation. To see this, consider how we might explain individual differences between people who are given a list of herbs to memorise. Suppose that some people know that 'rosemary' is the name of a herb, and others know it only as a name. Then, in this thought experiment, we might conjecture that the latter group would exhibit the von Restorff effect for 'rosemary,' because they would believe it to be the only non-herb in the list, and hence separate from all the other herbs. The former group would, presumably, show no such effect. Suppose this pattern of results were to occur. Then, as ever, personal-level explanation and cognitive science explain complementary aspects of the phenomenon. Personal-level explanation, along the lines we have described above (concerning the knowledge of the people in the experiment) determines whether the item is surprising; and then some scientific explanation might explain why particularly surprising items are particularly well-recalled (perhaps because they are encoded more elaborately (Craik & Lockhart, 1972), or are more distinctive (Nairne, Neath, Serra & Byun, 1997)).

Language comprehension

We can see the same pattern in research on language comprehension. For example, consider the problem of local ambiguity in language. As utterances are encountered word-by-word, they are typically highly ambiguous. For instance, *coach* has a different meaning in *The coach was very angry* from *The coach was very full.* But when *coach* is first encountered, the appropriate meaning is not apparent. A similar situation occurs for syntactic ambiguities. Thus, the fragment "The horse raced past the barn..." is typically initially interpreted so that 'raced' is an intransitive past-tense verb, whose subject is 'the horse.' But the completion "The horse raced past the barn fell" is inconsistent with this local interpretation. Instead the structure is a so-called reduced-relative, meaning "the horse, which was raced past the barn, fell." On hearing "The horse raced past the barn fell" people frequently get 'stuck' and falter in finding a correct global interpretation of the sentence (Bever, 1970). Not all such sentences produce such dramatic effects, but eye-tracking studies reveal people's widespread difficulties with syntactic ambiguities.

There are many accounts of how local syntactic ambiguity is resolved (e.g. MacDonald, Pearlmutter, & Seidenberg, 1994), but it is generally acknowledged that *plausibility* substantially affects which analysis is adopted. Now, plausibility is, of course, judged against the rest of background knowledge. Intuitively, the question is how well does the new information integrate with what is already believed; hence it falls squarely in the territory of personal-level explanation.

For example, Ferreira and Clifton (1986) and Trueswell, Tanenhaus, and Garnsey (1994) considered the processing of sentences like "The defendant examined by the lawyer turned out to be unreliable" and "The evidence examined by the lawyer turned out to be unreliable". In both cases, the correct reduced relative analysis of the sentence is pitted against a main clause analysis. In the first, this analysis is plausible, because defendants can examine things. In the second, it is implausible, because evidence cannot examine anything. Trueswell et al. found that people had difficulty reading "by the lawyer" after "defendant", and argued that they initially preferred the main clause analysis; but found that they had no difficulty reading "by the lawyer" after "evidence", and argued that they initially preferred the reduced relative analysis here. On their account, people made effectively instantaneous use of plausibility to determine which analysis to adopt. Although Ferreira and Clifton's evidence led them to the different conclusion that plausibility was not used during initial processing, they found effects of plausibility in subsequent processing. Both accounts found effects of plausibility on choice of analysis. Indeed, demonstrations of effects of plausibility on syntactic processing are widespread (Pickering & Traxler, 1998)

Does this mean that a scientific account of language processing–namely a specific psychological theory of the parsing process (which might, perhaps, be embodied in a computational model)–overlaps with the domain of personal-level explanation? It might appear to do so, because the psychological account makes central reference to the notion of the plausibility of an interpretation, in the light of background knowledge; and explicating what is, or is not, plausible in light of relevant background knowledge seems paradigmatically to be the territory of folk psychology. But, just as for memory, this appearance is misleading. The mechanistic psycholinguistic theory simply uses plausibility as an input, and then traces the ramifications of plausibility judgements for the process of language understanding. It is left to personal-level explanation to elucidate why a particular interpretation is more plausible than another–by specifying relevant knowledge about the domain. Again, the two styles of explanation are complementary–both are required to explain the performance of the language processor on some specific sentence; but they cover distinct and non-overlapping explanatory domains. For example, many experimental studies manipulate plausibility between conditions (e.g., Pickering & Traxler, 1998). A new set of participants (separate from those in the main experiment, but drawn from the same population) might rate a set of sentences for plausibility (e.g., on a scale of 1–7). One condition would consist of sentences judged very plausible, and one of sentences judged very implausible. But there is no attempt to provide any scientific account of *why* particular sentences are judged to be plausible.

It might be objected that the distinction between the domain of personal-level explanation and scientific analysis is, in this case, somewhat blurred. This is because there is now a range of techniques (Landauer & Dumais, 1997) which can automatically derive, from language corpora, judgements strikingly correlated with the output of personal-level explanations. Thus, a statistical analysis of language may mimic predictions from personal-level explanation. Crudely, statistical analysis might reveal that on hearing "the horse raced..." on previous occasions, the correct analysis was that 'raced' is an intransitive past tense verb; but it might be that for past occurrences of "the missile fired..." the correct analysis was a reduced relative.

If listeners are sensitive to such statistical properties, then their preferences for local ambiguities might be explained using purely statistical properties rather than personal-level explanation. This might suggest that personal-level explanation, which explains aspects of local ambiguity resolution in terms of the application of relevant knowledge, may overlap with mechanistic explanation, based on a computational analysis of statistical properties of language. But in reality this case show the possibility of *boundary disputes* between the two types of approach. Presently, it is unclear whether apparent effects of plausibility should be explained statistically (using a cognitive scientific explanation) or background knowledge (using personal level explanation). But, whichever viewpoint is correct, just one style of explanation will be appropriate—there will be no explanatory overlap.

Personal-level explanation and the brain sciences

The case of neuroscientific explanation is straightforward. With respect to *cognitive* processes, such as memory, language, and perception, neuroscientific explanation covers a very much narrower territory than cognitive science, and hence the non-overlap between neuroscience and folk psychology follows immediately. But overlap may seem more likely in relation to the neuro-scientific basis of, for example, abnormal mental states. Suppose a certain kind of thought disorder (e.g., schizophrenia) is caused by excess levels of a neuroscientifically identifiable factor (e.g., excess dopamine, Leonard, 1997). Now, a thought disorder itself is paradigmatically characterised in personal-level terms—the sufferer is conceived of as having bizarre *beliefs*. But the cause might be identifiable in purely biological terms.

Such cases are central to biological approaches to mental illness. Chemical imbalances, abnormalities in particular brain structures, and so on, are postulated as causes of mental illness, identified in personal-level terms; and treatment is then aimed at attacking the underlying biological cause. But notice that, as before, the personal-level and biological accounts tackle complementary and non-overlapping aspects of the phenomenon. The biological account attempts to identify an organic cause of the problem; but personal-level analysis characterises the problem itself (i.e., the specific pattern of beliefs, desires and actions the sufferer exhibits).

There is notoriously no overlap, and indeed, almost no linkage whatever, between these two types of explanation. Suppose that dopamine excess really does cause schizophrenia. How, then, does an excess of dopamine cause characteristic patterns of beliefs, desires and actions? There are essentially no serious hypotheses in answer this kind of question. From our perspective, this is not surprising: because to answer such a question would require neuroscientific explanation to extend into the territory of personal-level explanation; and this has generally proved infeasible.

Interestingly, the study of abnormal mental states is also subject to boundary disputes between personal-level and mechanistic styles of explanation, just as we saw in the study of cognitive processes such as language processing. Theorists differ concerning which mental illnesses originate from underlying biological factors (for which biological intervention, such as drug treatment, is the most natural form of therapy), and which arise from problems of life, and thought about life, which can most naturally themselves be understood and treated using personal-level explanation. Psychotherapy as a form of treatment typically assumes that disorders can be tackled at a personal level—therapeutic interventions involve conversational engagement with the sufferer to modify specific beliefs, desires, or patterns of thought; and the theoretical frameworks underlying psychotherapy are framed in personal-level terms.

Note, too, that, as in cognitive science explanation, so in the explanation of psychopathology, mechanistic and personal-level explanations may need to be interwoven. The personal-level might explain why particular life-events were viewed as traumatic; resulting negative thoughts might lead to persistently altered mood, with knock-on biological consequences for brain chemistry. Altered brain chemistry may then modify the formation of future beliefs and desires. Note that such patterns of explanation involve the interplay of different styles of explanation, rather than two levels of explanation of the same phenomenon. Neuroscience provides no explanation of specific sequences of thought and behavior, whether normal or disordered; it provides no explanation of people's beliefs or values (i.e., there are no neuroscientific facts, such as that an increase in dopamine, or damage to a specific brain region, leads to an increased liking for cats, or eliminates the belief that dogs are animals). That is, neuroscience does not encroach on personal-level explanation.

Rational choice explanation and folk psychology

Let us turn, now, to rational choice explanation, as applied in economics and, increasingly, others areas of social science (Elster, 1986). The approach can be motivated from simple assumptions concerning how it is reasonable to choose between sets of possible options (some of these options may have uncertain outcomes—and can be viewed as gambles).

An agent's choices can be modelled by a utility scale and a probability distribution as follows. Each (non-probabilistic) option can be assigned a utility value. And each probabilistic option can be assigned a value that is the *expected* value of its component non-probabilistic outcomes, where expected value is defined as the sum of the utilities of the non-probabilistic outcomes, each weighted by its probability. Here, the utility scale can be viewed as expressing the agent's preferences or 'desires'; and the probability distribution can be viewed as expressing agents' beliefs. We shall, however, return to the question of whether these identification with belief and desires is really appropriate. The machinery of rational choice can then be applied to people's observed choice behaviour, to infer their underlying utilities and probabilities.

Rational choice explanation can be used to explain a wide range of phenomena. For example, assuming that the economy is populated with rational agents of this kind can help explain the structure of supply and demand in the pricing of goods and services; the value that shares attain; and even, controversially, macroeconomic phenomena, such as the impact (or non-impact) of various attempts at government intervention in the economy. Moreover, it can be used to explain, at a broad level, why people choose their careers, how much of their time and money they invest in education, even, particularly controversially, whether they decide to marry and have children (Becker, 1975).

The approach can, moreover, be extended to explain aspects of animal behaviour. Optimal foraging theory assumes that animals distribute their foraging resources optimally, from a rational choice point of view—e. g., they leave a particular patch of food and search for another less depleted patch, at just the optimal moment (Stephens & Krebs, 1986). And conflict between animals (e. g., over mates) can also be usefully analyzed using the machinery of game theory (e.g., Maynard-Smith & Price, 1973).

Rational choice theory appears to represent a different kind of case where personal-level and scientific explanation overlap. As we have noted, the theory models people in terms of fundamental attributes concerning utility and probability; and it is natural to identify the scale of utility ascribed to a person as capturing desires; and the probability distribution ascribed to a person as capturing beliefs. On this reading, rational choice theory does indeed look like a scientific theory that immediately encroaches on the territory of personal-level explanation.

But this appearance is misleading, and it arises because rational choice theory can be treated in two completely distinct ways. One interpretation of rational choice theory takes seriously the identification of utility with desire, and probability with belief. From this standpoint, rational choice theory does indeed overlap with personal-level explanation; but on this interpretation rational choice theory is both false and excessively narrow. The theory is false, because there is an endless stream of empirical results that show that people do exhibit intransitive preferences, fall into all kinds of probabilistic and logical fallacies, fail to assign utilities to items in a consistent way, and do not assign utility in uncertain choice by the expectation of utility assigned to each determinate outcome (e. g., Evans, Newstead & Byrne, 1993; Kahneman & Tversky, 2000). So rational choice theory, considered as a set of principles governing how people's beliefs and desires interact can be decisively rejected. But more significant, perhaps, is the narrowness of the approach—rational choice is applicable only in extremely well-defined situations, where there are small numbers of outcomes. In everyday situations, by contrast, there are typically vast numbers of possible outcomes, none of which can be straightforwardly assigned a probability; and the utilities of those outcomes are also largely unknown. So, for example, in choosing a house, assessing the probabilities of various outcomes (an accident when emerging from a blind turning?) is enormously difficult; and judgements that concern utility are equally difficult. To prefigure later discussion, these problems are amplified because each individual judgement requires making many further judgements. To assess the risk from the blind turning, I have to consider the flow of traffic, the degree of visibility, the likely effectiveness of possible pre-emptive measures, and so on. The task of establishing all these probabilities and utilities seems endless; we shall argue below that it is, at least for any practical purpose.

To sum up, so far: we have considered the possibility that rational choice theory might be viewed as encroaching directly onto the territory of personal-level explanation, by providing a precise calculus of belief and desire. But we have seen that, on such an interpretation, rational choice theory is both empirically false, and also essentially inapplicable in most everyday choice situations.

If this were the only interpretation of rational choice theory available, then the entire approach would be fatally compromised. But there is another interpretation: which completely severs the connection between utility and the personal-level notion of desire; and any connection between probability and the personal-level notion of belief. According to this viewpoint, rational choice theory should be viewed as an instrument with which to describe the structure of the problems faced by agents. The key assumption of the theory is that the behaviour of such agents will tend, in the long run at least, to be 'sensible,' from the perspective of the description of the problem that has been given—i. e., this means that their behaviour should align, to some approximation, with the dictates of the rational choice model. But, from this perspective, this implies nothing at all about mental state of the agent, considered at the personal-level. We might call this the 'deflationary' conception of rational choice explanation.

The deflationary interpretation is patently that in play in applying the approach to animal behaviour (Maynard-Smith & Price, 1973; Stephens & Krebs, 1986). Behavioural ecologists using game theory to explain the conflict between two stags do not, of course, assume that their analysis relates to game-theoretic calculations within the heads of each stag (i. e., there is no assumption that one stag formulates beliefs about the beliefs and intentions of the other stag, and hence informs its own beliefs and intentions). Similarly, accounts that explain aspects of animal foraging in economically optimal terms do not require that the animals themselves are running through such calculations. Rather, in such cases, it is simply assumed that there is some mechanism that tends to amplify some behaviours at the expense of others, depending on their usefulness (this might be through natural selection, if the behaviour is genetically controlled; or through learning, if it is learned). Hence, 'rational' patterns of behaviour will tend to be propagated, and become dominant; other patterns of behaviour will tend to be extinguished. To drive the point home: rational choice explanations prove very useful even in explaining the behaviour of extremely simple animals, such as insects—to which the personal-level ascription of beliefs and desires is certainly unwarranted.

More importantly, the same deflationary interpretation appears to be equally appropriate for explaining human behaviour. For example, game-theoretic explanations of human economic behaviour notably do not assume that people go through game-theoretic reasoning—indeed, it is well-known that such reasoning is extremely difficult for people to conduct successfully (e. g., Colman, 1995). Indeed, even in carefully controlled laboratory studies, there is little relation between the reasoning processes that people go through and the logic of game-theoretic analysis. According to the direct interpretation of rational choice explanation that we discussed above, this evidence would count as further empirical refutation of game-theoretic explanation. But economists typically argue that game-theoretic analyses do not aim to model people's internal cognitive calculations. Instead, game-theory describes equilibria that people will be likely to find, perhaps more or less by trial and error, or at least through relatively simple learning processes, after they have played the game repeatedly. Thus, game-theory is viewed as describing the structure of the task environment, rather than saying anything concerning the beliefs and desires of the players. The same argument holds across other applications of rational choice explanation or, indeed, to rational explanation in the social, biological and cognitive sciences more generally.

Summary

We have argued, using a series of case studies, that despite appearances, scientific explanations of the mind, although diverse, consistently do not encroach upon personal-level territory. Thus, in practice, there appears to be no prospect of personal-level explanation being replaced by scientific analysis (as Churchland, 1986, might predict); or that science will assimilate and make rigorous personal-level explanation (as, e. g., Fodor & Pylyshyn, 1988 suggest). The question then arises: what is the origin of this explanatory divide? The next section tackles this question.

Why do personal-level and scientific explanation not overlap?

We have so far argued that there is no overlap between the territory of scientific and personal-level explanation. But why? According to many (e. g., Fodor & Pylyshyn, 1988), cognitive science is presumed to be formalized personal-level explanation. But, we argue, personal-level cannot *be* formalized (at least, presently), because common-sense knowledge, more generally, resists formalization. Hence, personal-level explanations cannot in practice be converted into scientific computational explanation. We argue that the non-formalizability of personal-level explanation is one of the most important lessons from traditional AI research.

As we noted above, research in expert systems and other aspects of 'good old-fashioned artificial intelligence' (Haugeland, 1989) can be viewed as attempt to build computational theories out of personal-level explanation: i.e., as attempting to formalize human knowledge, and use it as the foundation for a computational inference system. The research strategy was to understand everyday argument, as embodied in personal-level explanations, by specifying (i) the everyday knowledge involved, (in verbal form) and (ii) the mechanisms of inference used (e.g. inference in a particular logic; or, in relation to action, principles of rational choice theory, described above).

Unfortunately, so far it has proven to be impossible to carry out either aspect of this program successfully. Perhaps the most important reason is that everyday, common-sense knowledge appears not to decompose into separate domains that can be formalized independently. Indeed, common-sense knowledge appears to have an ineliminably open-ended character–in making or understanding any particular common-sense argument, there seems to be an indefinite, and arbitrarily disparate, body of knowledge involved (see Dreyfus & Dreyfus, 1986; Fodor, 1983).

To illustrate this point, let us consider John, who returns from work to find the lock on his front door broken, and infers that he has been burgled. This inference is intuitively straightforward. But the knowledge it involves is astonishingly rich, including facts such as: that burglars must enter a house before stealing from it; that a door is a potential entrance; that locked doors cannot be opened; that forcing the lock will allow the door to be opened; that locks do not spontaneously break; that people do not generally force locks without criminal intentions; and so on. But it also includes information concerning alternative explanations: for example, that John's wife, who might conceivably have locked herself out and had to break in, is currently in Greece; that Greece is too far away for a day trip home; that he phoned her there yesterday; she had no plans to return. There seems to be a never-ending stream of relevant information, which can come from entirely unexpected quarters (e.g. flight times from Greece). The experience in the AI community of building systems designed to carry out common sense inferences makes it painfully clear that if these pieces of information are ignored, inference is likely to go hopelessly awry. Fodor (1983) makes this point arguing that common-sense inference is always what he calls *isotropic*: relevant information can be drawn from anywhere and everywhere, and cannot be separated into evidentially disconnected domains.

A further, and devastating, difficulty is what has been termed the "fractal character of common-sense" (Chater & Oaksford, 1996): each argumentative link that we postulate is itself as complex as John's original inference. So, for example, the inference from John's speaking by phone to his wife the day before, to the fact that she cannot be back already, will involve endless information about airlines, alternative means of travel and their speed, and so on.

We have considered problems of capturing the knowledge involved in common-sense inference. Equally large problems arise in understanding common-sense inference. The most immediate problem is that common-sense inferences can be overturned by the addition of further information, i.e., they are non-monotonic. For example, John's inference that he has been burgled above may be overturned if he finds that his daughter has come back from university and has forgotten her key.

We illustrate the problems encountered in formally modelling non-monotonic reasoning by considering a popular and direct approach: attempting to develop a non-monotonic logic for everyday reasoning. For example, Reiter (1985) distinguishes between two kinds of information: certain, hard facts about the world, and default rules which support plausible inferences. The set or sets of beliefs sanctioned by a default theory are obtained by starting with the hard facts, and repeatedly applying the default rules to derive plausible, but not certain, conclusions. The intuition is that default inferences should be allowed when, but only when, their conclusion is consistent with what is already known. So, for example, a default rule from the premise that Tweety has wings to the conclusion that Tweety flies should apply unless the conclusion is not consistent with other known information (e.g. that Tweety is an ostrich).

Non-monotonic reasoning systems face a fundamental problem: of reconciling evidence that points to conflicting conclusions. Suppose that there is a second default rule, that badly injured creatures cannot fly, and consider what can be concluded from two hard facts: that Tweety has wings and is badly injured. In practice, it is clear that the latter rule should take priority–badly injured birds are unlikely to be able to fly. This follows because of what we know about the *meaning* of the terms involved, what wings, flight and injury refer to, and our general world knowledge about how these are related. But, of course, neither meaning nor general world knowledge can be used by computational accounts of non-monotonic inference, such as Reiter's, which depend only on the *form* of the predicates mentioned, and hence there is no way of resolving conflicting evidence sensibly.

It is therefore not surprising that Reiter's system cannot handle cases of conflicting defaults successfully. If the first default rule is considered first, the conclusion that Tweety flies can be drawn, and the second default rule is blocked, since its conclusion is inconsistent with what has already been derived. But if the second rule is applied first, the opposite conclusion that Tweety does not fly can be drawn, and then the first rule is blocked. All that can be concluded overall is that disjunction of these two conclusions holds: that Tweety either flies or does not. This "problem of weak conclusions" (McDermott, 1987) is endemic to formal approaches to non-monotonic inference: information about form is simply not sufficient to specify how conflict should be resolved.

We have so far viewed these difficulties as concerning the requirement of modelling common-sense inference. But equally, they bring out the problem that the relevant knowledge has not been encoded. In the above example of conflicting defaults, the system has not been given enough information to decide which is the most sensible way to resolve conflicting defaults; and formally encoding such information appears to be a literally endless task, since each new piece of knowledge will be as defeasible as the rest, and will require still further knowledge to specify how its defaults should be resolved. The open-ended character of common-sense, when squeezed into a logical representation (and equally a verbal representation), has suggested to some that such representations are not appropriate for representing common-sense knowledge (Dreyfus & Dreyfus, 1986). There is however a resounding silence concerning alternative forms of knowledge representation that might escape these difficulties.

How is personal-level-explanation possible?

We have argued that personal-level explanations are open-ended and draw upon an indefinite amount of knowledge; and that this is one of the reasons that it has not proved to be possible to model such inferences computationally. But how is it, then, that succinct personal-level explanations can be used in everyday life?

The answer is that in everyday personal-level explanation, only the bare bones of the argument need to be specified, since the rest, the indefinitely large body of knowledge which resists formalization, and the inference procedures over that knowledge, are common to the agent whose mental life is being explained and the audience to whom the explanation is addressed. Personal-level explanation involves attempting to use one's own mind to enter that of another. Hence, the knowledge and inferential machinery of one's own mind are given for free and need not be part of the explanation. In building a computational model, no such understanding can be taken for granted: the task in AI is to build a system which can make and understand such arguments completely automatically. And this is a task which, in view of the open-ended character of common-sense knowledge and the problems of capturing non-monotonic inference, has not been solved.

Conclusion

We have argued that the attempt to devise computational models based on formalized personal-level explanation has failed. We suggest that this personal-level explanation cannot readily be assimilated by science—there are no scientific accounts of what it is to have a particular belief or desire in information processing, or neuro-scientific terms. Hence, rather than serving as complementary levels of explanation or standing in competition, *in practice* personal-level and scientific explanations of the mind deal with non-overlapping aspects of mental life. Thus, there is, presently at least, no prospect that personal-level explanation will be either be assimilated to, or swept away, by developments in the neurosciences, or the cognitive and social sciences.

One implication, of particular relevance to this volume, concerns the relationship between belief ascription, as studied in philosophy, and the brain and cognitive sciences. If we are right, then there is little hope that difficult philosophical problems concerning belief and related notions will be resolved by appeal to cognitive science, because cognitive science simply does not engage successfully with those aspects of the mind underlying belief, desire and so on—i. e., as Fodor (1983) put it, there is no cognitive science of central cognitive processes.

This does not imply that philosophy and the brain and cognitive sciences should be pursued without reference to each other. Indeed we suspect that both explanatory approaches will be essential in elucidating thought and behaviour. Typically, cognitive activity engages *both* the elementary cognitive mechanisms of perception, motor control and memory *and* takes input from mental processes underlying belief, desire and common-sense inference. Thus, the interaction between the two realms may be of central importance in explaining almost any mental phenomenon of interest, whether primarily from a scientific or a philosophical point of view.

Nick Chater
Institute for Applied Cognitive Science
Department of Psychology
University of Warwick
Coventry, CV4 7AL,
U.K.
nick.chater@warwick.ac.uk

Martin J. Pickering
Department of Psychology
University of Edinburgh
7 George Square
Edinburgh EH8 9LW
U.K.
martin.pickering@ed.ac.uk