Chapter 4.

Metaphor, Self-Reflection, and the Nature of Mind

John A. Barnden
University of Birmingham, UK

Abstract

This chapter speculatively addresses the nature and effects of metaphorical views that a mind can intermittently use in thinking about itself and other minds, such as the view of mind as a physical space in which ideas have physical locations. Although such views are subjective, it is argued in this chapter that they are nevertheless part of the real nature of the conscious and unconscious mind. In particular, it is conjectured that if a mind entertains a particular (metaphorical) view at a given time, then this activity could of itself cause that mind to become more similar in the short term to how it is portrayed by the view. Hence, the views are, to an extent, self-fulfilling prophecies. In these ways, metaphorical self-reflection, even when distorting and inaccurate, is speculatively an important aspect of the true nature of mind. The chapter also outlines a theoretical approach and related implemented system (ATT-Meta) that were designed for the understanding of metaphorical discourse but that have principles that could be at the core of metaphorical self-reflection in people or future artificial agents.

Introduction: What Questions Are We Addressing?

(a) What is mind?

(b) What are theories of mind?

(c) What could or should computationally implemented architectures and systems based on theories of mind be like?

(d) How should we respond to a particular sort of fragmentation in the study of mind?

In this chapter, these questions are asked from the indirect point of view of *how a mind views itself or other minds* rather than directly from the theoretical observer's point of view of determining what the mind *really* is. As a result, reflection on the above issues (a through d) is roughly as follows, where (a) and (b) have been collapsed together:

(ab2) How does a mind view itself (what kinds of theories does it have about itself); how does it view other minds; and how do these matters interact with the question of what minds really are?

(c2) What could or should computationally implemented architectures and systems involving minds' views of minds be like?

(d2) How should we respond to a particular sort of fragmentation in minds' views of minds?

The move to these issues from Issues a through d might be thought to be twisting the latter too far. But behind *ab*2 is a claim that *how a mind views itself is part of, and can affect, the real nature of mind itself*. After all, one important aspect of mind is its process of thinking (consciously or unconsciously) about itself. *How* it views itself is then a bald fact about that mind. For instance, to make the point vividly, it may view itself as being a physical being trapped inside the body and able to have a life outside that body if only it could get out. Theories of mind must take into account the views and theories that minds have about themselves and each other, even if they are highly inaccurate or irrational. If a mind thought it was made out of fire and water, then the fact that it thought that is an important fact about that mind, even though it is not actually made out of fire and water.

But, more deeply, the author will claim that a view that a mind has of its own nature at some particular time can, so to speak, entrain that mind to become, at

around that time, more in accordance with that view than it would have been without the operation of the view. Views of oneself can, to some extent, become self-fulfilling prophecies.

The main ideas in this chapter concern an individual mind's intermittent views of itself, rather than of other minds, and is thus consonant with the need to include self-reflection capabilities in architectures of complete minds. However, a view that a mind has of itself at some moment could be influenced by the views it perceives other minds as intermittently having of themselves and each other. Views of mind, as with views of anything else, can be transmitted from person to person.

Issue (d), as opposed to (d2), is about disciplinary fragmentation in research on mind. The findings in this chapter implicitly contribute to curing (d) to some extent, even though it discusses (d2) instead, and is indeed in the direction of supporting the idea that the mind has a natural tendency to have a fragmented overall view of itself. The chapter material is disciplinarily integrative in bringing a computational outlook to bear on deep philosophical and psychological issues and naturally supports a link between the study of language about the mind and the study of mind; in particular, metaphor is given a central place in the study of mind. Moreover, "mind" is taken to include affect. The way a mind views its own affect is important here, as are the affective aspects of the way a mind views itself (even the nonaffective features of itself).

The plan of the chapter is as follows. Section 2 provides background on metaphor and its relationship to thought and the study of thought. Section 3 is the start of the main thrust of the chapter and argues that for reasons of practical necessity, self-reflection is likely to be importantly metaphorical. Section 4 discusses ways in which metaphoricity of self-reflection could distort the true nature of the mind as opposed to merely presenting a distorted picture of a mind to itself. Section 5 presents a case for some qualia in consciousness to be intrinsically metaphorical in nature. Section 6 discusses the fragmentary nature of self-reflection that is likely to arise from metaphor, but which may be unavoidable in any case. Section 7 briefly outlines the author's approach to the understanding of metaphorical discourse, and an implemented system (ATT-Meta) derived from it, and shows how the approach could provide principles and techniques that could be at the core of a metaphorically self-reflective agent. Section 8 concludes.

Throughout the chapter, it is important for the reader to bear in mind that the main concern is with what views a mind might take of its own internal states and processes in the short term for some particular cognitive purpose, rather than with views a mind might continuously have about itself in the long term. Of least concern is the question of how similar those views are, or should be, to attempted objective accounts of the mind that might be devised by scientists or philosophers:

in other words, accounts that minds might have of themselves as a result of extended intellectual deliberation as opposed to resulting from the ordinary life experience that anyone could have without being a scientist or philosopher.

Background: Metaphor, Discourse, Mind, and Affect

One aspect of a complete mind, situated within anything like our world, must be the ability to reason about other minds and about itself as a complete mind. Now, as cognitive linguists and others have shown (see, for example, Lakoff, 1993), much human discourse concerning minds is highly metaphorical. Some particular, common ways in which discourse talks metaphorically about mind are as follows.

- **Mind as Physical Space.** We commonly talk about minds as if they were physical containers or physical regions. This is typified by utterances such as "The idea invaded my mind," "She pushed the idea to the back of her mind," "The fear was buried deep within his mind," and "In the far reaches of her mind, she knew that her husband had been unfaithful."
- **Ideas as Living Creatures.** We commonly talk of ideas as if they were living creatures, as in "The belief had been lurking in her mind," "Several different desires were battling inside her," "The thought of his impending arrival was tugging insistently at her," and one of the examples above of *Mind as Physical Space*: "The idea invaded my mind."
- **Ideas as Physical Objects.** The metaphorical view of *Ideas as Living Creatures* is a special case of a more general, pervasive metaphorical view of ideas as physical objects that could be inert. This appears in utterances such as two of the examples above of *Mind as Physical Space:* "She pushed the idea to the back of her mind" and "The fear was buried deep within his mind;" as well as in utterances such as "The suspicion stuck to her like a magnet," "They kicked ideas around the room," and "She was still a long way from a solution," where the ideas are in an external physical space (either a real one surrounding the person mentioned or an imaginary physical space outside the person).
- **Cognition as Perception.** We often talk of cognition as if it were an included physical perception, as in "She couldn't focus clearly on the problem," "He could picture it very clearly in his mind," "The memory was

lost in the mists of her mind," and "The situation stank of corruption." The visual case, as in the first three of these examples, is particularly common.

- **Mind Parts as Persons.** A metaphorical view that is less often remarked upon, but that is nevertheless commonplace, and plays a significant role in this chapter, is where a mind is viewed as being made up of or containing subpeople, with their own thoughts, desires, etc., and possibly communicating with each other, as in "Part of him was afraid of raising the issue," "Part of me could see that my sister had been dishonest," "One part of me was whispering that I ought to leave, while another part was begging me to stay," and "The child inside him was crying for attention." Locutions such as "being in two minds about *X*" should perhaps be classified here as well, together with some of the metaphors discussed by Lakoff (1996).

Other metaphorical views of mind, together with abundant examples, can be found in many sources, including the author's own databank at http://www.cs.bham.ac.uk/~jab/ATT-Meta/Databank, which also has links to other sites. (We use the term "metaphorical view" to mean roughly what Lakoff and others mean by a "conceptual metaphor"—essentially, a mapping from aspects of one domain to supposedly corresponding aspects of another.)

Discourse can switch rapidly between different, possibly conflicting views in describing mental states and processes. Discourse often mixes different views together, as will be illustrated below and as is already exemplified by the fact that some of the utterance examples listed above are included under more than one view. This is all in perfectly mundane discourse, not (just) poetry and other literary art, as our examples illustrate.

A further tenet held by many metaphor researchers (for example, Gibbs, 1994; Lakoff 1993) is that the metaphorical views used in discourse are, generally speaking, crucial aids in thought (conscious or unconscious) rather than just linguistic decoration. In this chapter, this is assumed to be true. Thus, our starting point is that people's conscious and unconscious thinking, not just their discourse, about each other and about themselves is partly and perhaps highly metaphorical. This is not to say that minds *believe* that the views are true, they are just useful ways of thinking, consciously or unconsciously. We also assume that people can adopt different views at different times and in different circumstances, and can switch rapidly between different views, even when thinking about one person's thoughts, just as in discourse.

The fact that people use metaphorical views not only in thinking about other people's minds but also in thinking about their own is suggested by the observation that metaphorical talk about mental states are often in the first person. In the case of *Mind Parts as Persons*, for instance, spoken uses of the

view are often, and perhaps mostly, in the first person, although third-person uses are common in novels, etc. As for other metaphorical views, examples such as the following are common in ordinary discourse:

- It was in the back of my mind.
- The thought crept into my mind.
- The thought stuck to me.
- I said to myself that ...
- My mind felt totally focused.

We will assume in this chapter that there are several possible ways in which a particular person can be led to use a particular metaphorical view in thought or language. People may, possibly, be genetically predisposed to think of minds in a particular way. They may develop particular ways as a result of observing the ways their own minds work and considering how other people's minds may be working. Or, they may learn to use a view because they encounter it frequently in discourse. Clearly, therefore, the views a particular person uses may be strongly influenced by the language and culture that person is embedded in and by societal norms laid down in that culture about how people should think of themselves or others (see, for example, Johnson, 1985). Another cultural aspect is that much of religious belief is about the nature of the self. The degree of variety, therefore, in a particular person's use of metaphor of mind may be partly dependent on culture; however, an added complication is that the way metaphor enters into the person's unconscious thoughts may be different from the way it enters into their conscious thoughts, and from the point of view of the present discussion, unconscious thoughts are of great potential importance.

In natural language discourse, affective states are often described metaphorically (see, for example, Fainsilber & Ortony, 1987; Kövecses, 2000). Examples are "His anger boiled over" and "Sweet feelings welled up within him." A mind's internal reflection on its own affect can therefore be conjectured to involve metaphor. Also, metaphor is often used in natural language discourse to convey value judgments and emotions about the targeted subject matter (often to deviously smuggle them in, but also, more beneficently, to convey them in an economical and effective way). For instance, hearing poverty being described as a disease could cause one to have particular negative emotions about poverty or poor people. Similarly, thinking of poverty as a disease could have such effects. Thus, we may conjecture that a mind's affective states can, in part, be caused by that mind's metaphorical thoughts as well as being explicitly described in that mind's metaphorical thoughts.

We assume also that a metaphorical view that someone takes of some entity can affect not just the person's reasoning, emotions, value judgments, and communication about it but, therefore, also how the person deals with it (interacts with it, manipulates it, controls it, etc.). For example, thinking of poverty as a disease can affect one's attempts to control it or can affect one's interactions with poor people. This effect is at the root of the use of metaphor in political discourse in order to persuade people to adopt particular stances or behaviors (see, for example, Mio, 1997).

Discussions of self-reflection, except when conducted by a metaphor researcher, rarely engage with the way that metaphor might enter into self-reflection. This is one example of the fragmentation of research emphasized by Issue (d). We must note carefully that many discussions of consciousness refer to or use metaphorical notions of mind, such as the Cartesian Theatre and internal narratives (Dennett, 1991), the mind's eye (Rorty, 1980), and global workspaces (Baars, 1993). However, what is predominantly at issue here is the question of what metaphors it is appropriate for an observing theoretician to use or to avoid in elucidating the true nature of consciousness, *not* with the question of how the use of such metaphors within the self-reflection of the observed mind can affect the nature and functioning of that mind. One type of exception to this neglect is the study within the area of psychiatric therapy of how people's metaphors about their own selves affect their moods and their thoughts about themselves (Mio & Katz, 1996). Of course, the popular self-help literature is full of concepts such as "inner child" and "playing negative tapes in one's head," together with instructions about how to attend to, exploit, or avoid such metaphors in controlling one's inner and outer lives, although the metaphoricity may not be explicitly recognized in a particular tract.

We take the views discussed in this section to be metaphorical, even though there may be senses in which some of them could be construed as literal. For instance, if one holds the philosophical view that ideas are neural patterns of activation and allows that such a pattern is a physical object, then ideas really are physical objects. They would then also have particular physical locations inside the head, whether in the sense of being physically confined to one small region of the brain or in the more general sense of spread over many widely distributed, but nevertheless specific, neurons in the brain. Furthermore, the ideas could change location. However, we are concerned with the *Mind as a Physical Object* view as part of the common sense of an ordinary person, and as part of philosophically and scientifically uninformed discourse, not dependent on knowledge of how the brain works or of theories about the relationship between mind and body. Another example is that one might propose that the mind is in fact made up of subagents that possess separate beliefs, desires, etc. However, even if it is, discourse involving *Mind Parts as Persons* is not dependent on it being the case or on any suspicion by the participants that it is the case, and that metaphorical

view may conflict with the discourse participants' scientific theories, if any, about the nature of mind. A further observation about *Mind Parts as Persons* is that while it is sometimes used with an implication that the mind parts in question are permanent features of that mind, as in "My inner child is always feeling jealous so my adult self spends a lot of time arguing with it," it is also often used with no such implication, as in "One part of me can see the force of that argument," where the part is not further characterized in the discourse as having any special, long-term existence or qualities.

Metaphorical Self-Reflection as a Practical Necessity

In the previous section, metaphorical self-reflection was mentioned as a mere special case of self-reflection in general. However, it is plausible to suggest that there may, in fact, be no practical alternative to doing substantial amounts of reasoning about mental states by means of metaphor.

First, as can be seen from the literature on metaphor, such as the work by Lakoff (1993), it is plausible that there is no practical alternative to metaphor for thinking about messy abstract domains, especially when matters are complex or subtle. It is widely acknowledged that metaphor, when applied appropriately to messy domains, can provide more economical and precise description, and more effective reasoning, than is otherwise practical. It may not even be possible to describe some things without metaphor (see, for example, Stern, 2000) on the problem of unparaphrasability of much metaphor. As an example of the difficulty of doing away with metaphorical description, it is difficult to paraphrase the sentence "In the murky depths of her mind, Anne realized that her husband had been unfaithful" without resorting to other metaphors that capture ways in which Anne's thoughts can be relatively inaccessible to her. At least, it is difficult to do without resorting to lengthy circumlocution.

Second, it is plausible that one's own mind is a messy, complex, yet subtle domain for oneself as well as for others, even if one has some sort of privileged, direct access to one's own mind. The potential messiness, complexity, and subtlety is, if anything, increased by having more extensive access to one's own mental states than to those of other minds.

Furthermore, an agent *X* can be expected to learn metaphorical ways of talking and thinking about minds, from the metaphorical ways that other agents use in their speech. These ways of talking about minds could be absorbed by *X* and become ways in which *X* thinks about itself. This does not, of course, preclude

X developing metaphorical and other ways of thinking about itself purely through self-reflection.

Distorting Oneself Through Metaphor

We pointed out above that a metaphorical view used in a mind's self-reflection is a real feature of that mind, no matter how inaccurate the view is. Metaphorical self-views are aspects of the real nature of mind. But, we can also see ways in which the use of a metaphorical self-view at some time can cause a mind to become, at around that time, more similar to its own view of itself than it would otherwise have been. In other words, views of oneself can become self-fulfilling prophecies, to some degree. We will now look at two possible ways in which this could happen.

Distortion Method 1: Through Metaphorical Self-Management

Any given broad metaphorical view of mental states and processes, such as *Mind as Physical Space* or *Ideas as Physical Objects* captures some real aspects of mind and ignores others. This is just a special case of a general feature of metaphorical description—different metaphorical views generally capture different aspects of what is being viewed (Grady, 1997; Lakoff & Johnson, 1980). Moreover, even when a view captures certain features, it generally does so only approximately. For example, a view of marriage as a journey approximately captures ways in which the relationship can develop, whereas a view of marriage as a business contract approximately captures ways in which the partners interact or should interact with each other at any moment in time.

Thus, if a mind's self-management is partially influenced at some point in time by a particular metaphorical view *V*, the self-management may be partially defective because of the inaccuracies of *V*. But, the operations of self-management may, to some extent, tend to make the mind behave as if it were more accurately described by *V*. We can call this phenomenon distortion of oneself through metaphorical self-management.

A vivid commonplace example of this potential effect is when people view themselves as having an "inner child" partially governing their thoughts and actions. To the extent that they believe that children should not be overly controlled, or cannot be controlled, they may refrain from taking self-control

actions that they would otherwise take (and be perfectly able to take). Similarly, to the extent that they believe a child should be controlled or influenced by certain methods, they may take analogous self-control actions. In these ways, they would come to act and think more as if they really had a child inside controlling things.

One general point here is that if a metaphorical view used in a mind's self-reflection fails to be sensitive to particular opportunities for external or internal actions by that mind, self-management may be deprived of the opportunity for exploiting those possibilities, so that the mind does not perform actions that it could, in fact, perform. To take a less vivid example than that of the inner-child, a certain person *Z* may at some point in time be viewing her own current mental operations via the metaphorical view of *Mind as Physical Space*. Some of her self-management could be influenced by her perceptions of how "central" some ideas are in her mind-space (that is, her perceptions of how strongly she is attending to them), and could be conducted with the aim of "moving" ideas closer or further from the centre. She might assume that bigger "movements" require more effort, so that less central ideas require more time and effort. She might therefore attend even less to ideas she perceives as being on the periphery of her mind, even if, unbeknownst to her, they do not require significantly more time or effort to deal with. Those ideas could then become less attended to than they were already. Thus, the metaphorical view *Z* she is taking of herself can tend to cause her to exaggerate certain features of her mental state that are (inaccurately) captured by the view.

This example is related to the more general observation that people can become limited by the views that they hold about themselves. Someone who believes they cannot do something will tend not to try to do it, whatever it is, and may also engage in behavior precluding it. Through such effects, the person may become less able in some capacity than they otherwise would have been.

The example above was about *Z* refraining from certain mental acts. More positively, a metaphorical self-view could lead *Z* to engage in certain types of mental behavior that she would otherwise have been less likely to engage in. For example, consider the common metaphorical view of *Ideas as Internal Utterances*, used in sentences like "One part of her whispered that she was wrong." Suppose that at some moment in time *Z* is experiencing and thus viewing some of her own thoughts as internal utterances. This may lead her to respond to those thoughts by further acts of mental utterance, when she might otherwise have responded by, say, constructing a mental picture or diagram. Thus, the view that happened to be in play has encouraged its own strengthening and continuance (for some period of time, which may be very short).

Distortion Method 2: Through Metaphorical Self-Conformity

The previous subsection can be roughly (and metaphorically) summed as follows: a self-reflective Part A of a mind views a Part B metaphorically and thereby distorts B (for example, by exaggerating certain qualities of it). The present subsection is about a Part B distorting itself by conforming to some view that another Part A is using in its actions toward B.

To some extent, people tend to adapt to—in the sense of coming to conform to—views that other people have of them or ways other people have of dealing with them. For example, if Person B thinks A thinks B is stupid, B can start to act more stupidly than he would otherwise. Another case is illustrated by a situation where a Person B acts in a businesslike way in dealings with A, because A is acting in a businesslike way with B. Quite apart from such questions as B's thinking that he *should* act in a businesslike way, in order, say, to impress or outwit A, there is simply the effect that A's businesslike dealing with B sets up a context of action where some types of action are more appropriate than others, and B may simply slide naturally into that context through a type of imitation.

Similarly, given that metaphorical views that people entertain about each other can affect the way they behave toward each other, it follows that someone may tend to conform, temporarily at least, to some view that is affecting the way someone else is dealing with him. B may find that A is viewing B's argumentation as physical attack and is, therefore, engaging in conversational maneuvers modelled on combat situations, thereby leading B to act similarly.

Now, could this type of metaphor-conformity effect apply also within a single mind? Suppose a mind can sometimes be legitimately viewed as being composed of two or more subagents, with mind-like capabilities and the ability to perform operations on each other, to reason about each other and to communicate with each other. That is, suppose that to some degree the mind really is, perhaps intermittently, organized according to the view of *Mind Parts as Persons*. (Even if the human mind is not normally, or ever, like this, it could be the way an artificial agent is organized.) Then is it too fanciful to suppose that a subagent B could conform to the way it is being dealt with by another subagent A, and in particular, come to behave more in accordance with some metaphorical view that A is entertaining about B? Note here that previous parts of this chapter lead to the conjecture that subagents would engage in metaphorical thought about each other just as much as a single unitary mind would engage in metaphorical self-reflection. For example, subagents might view each other as engaged in physical combat.

If such a metaphor-conformity effect could happen, it would be another way in which the overall agent is distorting itself to conform to its own metaphorical self-reflection.

Metaphorical Qualia

Suppose Bill works at the Foreign Office within the government of a hypothetical country and is metaphorically viewing the Office as a solar system, with the Foreign Secretary as the sun and junior ministers as planets. Suppose even that Bill is one of the planets. Then it could perfectly well be that Bill does not *feel* like a planet going round a sun, even partially. For example, he may well have no feeling of being, literally, physically pulled toward the Foreign Secretary, physically circling round him or her, or receiving life-giving radiation from him or her. It could well be that Bill has worked out, or learned from others, that there is a formal correspondence between certain abstract relationships and activities within the Office (or within organizations of that type in general) and the relationships and processes in a solar system.

However, it is possible that Bill has some feelings that are similar to the putative feelings mentioned in the previous paragraph. For example, it could be that the feeling of loyalty toward the Foreign Secretary has something in common with feeling physically pulled toward him or her, and, more strongly, that feelings of pleasure and comfort arising from being in his or her good books could be similar to, if not actually the same as, some of the feelings of pleasure or comfort arising from basking in sunlight. Such possibilities would fit well with theories that metaphor derives partially from embodied experience (see, for example, Johnson, 1987). Equally, as many commentators have observed, talk of being hot with anger may derive in part from feelings of being hot when angry. Thus, at least in the case of some metaphors, the conscious qualia involved in the target-domain situation being described (Foreign Office, say) may be, in part, similar to or the same as qualia involved in the source domain.

When applied to metaphors of mind, such considerations lead to an important additional line of thought that extends the suggestions in previous sections of this chapter. It is developed more extensively here than in Barnden (1997). We start with the observation made in an earlier section that first-person manifestations of metaphors of mind are common. The central conjecture of the present section is that we do not use such language for practical convenience, i.e., merely because it supports useful reasoning about the described mental states, but also because, at least to a limited extent and for some of the time, it reflects the *feel*

of mental states to us. Here the word "feel" has a sense as broad as the word "qualia." (Thus, in the intended broad sense, redness has a feel.) Using this broad sense, the conjecture is that, for instance:

- Thought can (sometimes) *feel* like internal speech.
- Thought can (sometimes) *feel* like vision.
- One's mind can (sometimes) *feel* like a physical space, and one can *feel* that one's ideas are far apart or moving around within that space, or coming into the space from outside.
- One may (sometimes) *feel* that inside of one there are several independent thinking entities with their own thoughts and feelings.

Now, if this is the case, then, because these feelings are part of the conscious mind, it follows that at least some metaphorical views of mind are, in part, intrinsic aspects of the nature of the consciousness, not just convenient tools for describing mental states. This is a new way, going beyond the points made in previous sections, in which metaphor-based self-reflection is part of the *actual* nature of mind and not just a matter of inaccurate views of the actual nature of mind.

The case of cognition feeling like vision and other types of perception is especially interesting, as it connects to the study of mental imagery in psychology and philosophy and to old debates about the role of imagery in cognition and consciousness (see Glasgow, 1993, for a review and an artificial intelligence model). To liken, say, conscious visual imagery to an activity of picturing is to say that conscious visual imagery feels, to some extent, like seeing a picture. Notice that this feeling is a real part of the person's current state of consciousness, even if it is epiphenomenal in the sense of it not having any effect on the person's mental processes. Also, see Horne (1993) on the sensuous nature of imagery.

One recent theory that relates cognition strongly to perception is that of Barsalou (1999). Barsalou downplays the role of metaphor, but the mental use of perception-based metaphors for abstract concepts, including concepts about the mind, is not antithetical to his claims. His theory, in relying heavily on simulation of perceptual processes and on the stimulation of related affective states, would then tend to support the idea that thoughts couched in terms of the source domain (for example, the domain of physical objects) in a metaphor for mind would make the person experience the qualia that would arise in that source domain.

Fragmentation of a Mind's Overall View of Itself

As we stated above, a given metaphorical view captures only part of the targeted phenomenon, and different views capture different parts (in general). Also, because of inaccuracies in the capturing performed by different views, the views can conflict in what they convey about the target. A business metaphor for marriage emphasizes competition, whereas a journey metaphor emphasizes cooperation.

Thus, it is natural to expect that, on the assumption that self-reflection in minds is importantly metaphorical, there will necessarily be an important degree of fragmentation and inconsistency in self-reflection, and these effects stand to be heightened by the potentially self-fulfilling nature of metaphorical views. This may sound like a disadvantage. However, given a need for self-views to be importantly metaphorical, it is an advantage for there to be multiple views: they can potentially offset each other or be applicable in different situations, providing, overall, a higher degree of accuracy and completeness of self-reflection than would arise from using any individual view.

The mentioned fragmentation and inconsistency are primarily an observation about human minds. However, metaphor could provide to artificial minds a useful tool for description of mental states, capturing complexities, subtleties, and messiness that it would otherwise be difficult to deal with. Thus, metaphorical self-reflection and management could be useful. But this would bring in fragmentation and inconsistency. This chapter proposes that this outcome should simply be embraced. After all, nonmetaphorical self-reflection would probably have to involve oversimplifications and, therefore, inaccuracies, and there might need to be multiple, partially inconsistent self-views even if they were all nonmetaphorical.

Toward the Future: The ATT-Meta System and Approach

The author developed a theoretical approach and an implemented artificial intelligence system, called ATT-Meta, for conducting metaphor-based reasoning (Barnden, 1998, 2001; Barnden et al., 1994; Lee & Barnden, 2001). This has been applied largely to the special case of metaphor-based reasoning about mental states. For example, it can trace through implications of two ideas being

"far apart" in a mind considered as a physical region. The intended ultimate purpose of the methods used in the system is for them to form part of natural language discourse processing. However, the techniques used in the system could also be used reflectively by a mind to reason about itself on the basis of metaphorical self-reflection. In this section, we comment on some features of the system and the underlying theoretical approach, in the spirit of indicating how the types of mental processing discussed in previous sections could realistically form part of a mind design.

In the ATT-Meta approach, the understanding agent (or an agent using metaphor in its own thought processes) is assumed already to have acquired knowledge of a range of commonly used metaphorical views. Recall that a metaphorical view is essentially a mapping of aspects of the source domain (for example, physical space) to aspects of the target domain (for example, mind). We assume that the individual mapping relationships making up the mapping are general in nature. For example, the ATT-Meta system's knowledge of the *Ideas as Physical Objects* view is largely a matter of a mapping relationship that maps physical manipulation (of ideas that are being viewed as physical objects) to mental usage of those ideas, with no specific types of manipulation, such as banging or sawing, being mapped. We also assume that the view maps physical interaction between an agent's ideas to conjoint mental usage of those ideas by the agent. Similarly, the *Mind as Physical Space* view is largely a matter of mapping physical presence in the space to existence in the mind in question.

The reason that such mapping relationships are powerful is that the approach allows for an indefinite amount of reasoning within the terms of the source domain. For instance, suppose an utterance says that two ideas are "far apart" in someone's mind. We assume the understander takes this to be portraying the ideas as physical objects that are physically far apart in that mind conceived of as a physical space (so the utterance relies both on *Ideas as Physical Objects* and on *Mind as Physical Space*). Given the common-sense source-domain knowledge that physical objects do not normally interact to any substantial degree when far apart (at least in the everyday physical world), we get the source-domain inference that the two mentioned ideas are probably not physically interacting to any substantial degree, and therefore, via the mapping relationship mentioned above, we get the target-domain conclusion that the ideas are probably not being used conjointly to any substantial degree by the agent in question (so, for example, the agent will not infer consequences of the two ideas taken together).

The source-domain reasoning can be much richer than this. For instance, if an idea is being portrayed as being in the murky depths of someone's mind, knowledge of how murk and physical depth can affect physical visibility, accessibility, and manipulability will be used to connect to general mapping

relationships of the type illustrated above. (In this example, the view of *Cognizing as Seeing* comes into play as well as the two views used above.) A distinctive feature of our approach to metaphor compared to most others is its allowance for an indefinite amount of complex source-domain reasoning; exceptions include the approaches of Hobbs (1990) and Narayanan (1997). Unique to ATT-Meta, however, is an implemented, tested system that allows the source-domain reasoning to be arbitrarily interleaved with other reasoning operations, such as target-domain reasoning and mapping operations between source and target. It is to be expected that this interleaving would be important for a realistic application of the approach to the self-reflection concerns of the present discussion.

The ATT-Meta approach has as one of its basic principles that of *Map-Extension Minimization*. This can be illustrated with the murky-depths example. The approach avoids, if it can, trying to map the murky depths over to the target domain. Rather, it is only the mappable *effects* in source-domain terms that are mapped over, in this case, the effects of being inaccessible, etc. The reason for adopting this stance is that in many cases there is simply not enough nonmetaphorical knowledge about how minds work to be able to find target-domain correspondents for things like murky depths, and we suspect that such concepts are used in metaphorical utterances purely for the effects they have. Another reason for the stance is that it can be extremely expensive in computational terms to search for a coherent partial isomorphism between two domains (Falkenhainer et al., 1989). In any case, we would claim that in many cases there simply is no isomorphism to be found or intended by the speaker.

Also distinctive in the approach is a worked-out and implemented account of how the compounding of metaphorical views works. This is hinted at in the above examples, as there were two or three views compounded. Our approach to compounding (which we also call mixing, though without the negative connotation that the term "mixed metaphor" is often taken to have) is discussed further in Lee and Barnden (2001). One feature of our approach is attention to the distinction between parallel mixing (where a domain is viewed at the same time in terms of several source domains, as in the above examples) and serial mixing (usually called chaining, where Domain A is viewed in terms of Domain B, which is in turn viewed in terms of a Domain C). A (real) example of serial mixing is "The thought of my mother-in-law's arrival hung over me like an angry cloud," where the thought is viewed as a cloud, and the cloud is viewed as an agent that has an emotion. Given that metaphor compounding is not rare in ordinary discourse about mind, it is reasonable to take it as something that would need to be accounted for in metaphorical self-reflection.

The approach and system allow for graded effects in two senses. First, there is a handling of uncertainty and of conflict between lines of reasoning. This is

important in view of the fact that common-sense knowledge and reasoning are typically uncertain (for example, it is only usually the case that physical objects that are far apart do not interact very much), and the outputs of metaphorical mapping may conflict with target-domain knowledge and reasoning, which may be uncertain. The other graded effect is that things can be the case to varying degrees: for example, objects that are close together can interact to a high degree, whereas objects that are far apart interact only to a low degree. Notice that such degrees are orthogonal to the question of uncertainty. Some low degree of physical interaction may be supposed to exist with high certainty, and a high degree of interaction may be supposed with low certainty.

As regards the nature of source-target mapping, we have so far only mentioned mapping relationships that are specific to particular metaphorical views, such as the mapping relationship from physical manipulation to mental usage. However, our approach (Barnden & Lee, 2001) also incorporates view-neutral mapping adjuncts (VNMAs), which are mapping principles that apply by default whatever the particular metaphorical views are in operation. For example, the temporal order of events in the source domain is assumed to map to give the same ordering of any corresponding events in the target domain. Another example of a VNMA is that the degree of ease with which something can be done in the source domain maps to the degree of ease of corresponding actions (if any) in the target domain. As illustrated in Barnden (2001b), it is often the case that many of the important effects of a metaphorical utterance occur via VNMAs rather than by the view-specific mapping relationships, which merely supply a substrate of correspondence on which VNMAs may then work. A few VNMAs have been realized in our implemented system, but considerable work remains to be done on implementing them.

From the point of view of the present discussion, a particularly important VNMA is one that transfers information about emotion from source to target. Specifically, if in the source domain an agent has an emotion about a situation, and both the agent and situation map over to an agent and situation in the target domain, then (by default) the target-domain agent is assumed to have the same emotion about the target-domain situation. Thus, if a mind is self-reflectively reasoning about its own emotions in a metaphorical way, then it can create useful results about its own emotions. Another aspect of the VNMA is that the metaphorically reasoning agent's own emotions about a source-domain situation are also assumed by default to be emotions about target-domain situations (if any). For example, if the self-reflecting mind is saddened by something expressed in source-domain terms, then it is (probably) saddened by the real situation portrayed.

As explained in Barnden, Glasbey, and Wallington (to appear) and Barnden, Glasbey, Lee, and Wallington (2004), the ATT-Meta approach makes some

major claims about discourse-extended metaphor. The approach claims that it is a mistake to assume that the metaphor side of the task of understanding discourse is to convert each metaphorical utterance into nonmetaphorical internal-representational terms. It is often better, more economical and more effective, for the understander to keep thinking in terms of the source domain over the course of understanding several utterances, and only cross over to the target domain when there is a real need (for example, when information needs to be integrated with information conveyed by nonmetaphorical utterances). In this way, some individual utterances may merely contribute to an overall source-domain scenario out of which selected aspects are mapped (aspects that may have no simple relationship to any one utterance), rather than contribute target-domain information by themselves. We should expect that in metaphorical self-reflection it could be useful to pursue the self-reflection over a considerable period of time in source-domain terms, only crossing over into the target domain when necessary. In short, not every episode of metaphorical self-reflection need have direct target-domain consequences of its own.

Closely related to these ideas is our argument (Barnden, Glasbey, Lee, & Wallington, in press) that it is useful to be able to map information from target to source as well as in the usual direction of source to target. For example, when a metaphorical view is extended over a stretch of discourse, information derived from interspersed nonmetaphorical utterances can beneficially be converted into the terms of the prevailing metaphorical view(s). This allows integration of information to happen in source-domain terms rather than in target-domains terms. We argue that integration on the source side is often easier and richer than integration on the target side. Correspondingly, in metaphorical self-reflection, it could be beneficial to engage in target-to-source mapping to achieve integrated thinking about oneself.

Finally, the ATT-Meta system has facilities for reasoning uncertainly in nonmetaphorical terms about agents' beliefs and reasoning. It allows for any degree of nesting (reasoning about agents reasoning about agents reasoning about …). Thus, the system is relevant to arbitrarily nested self-reflection. Metaphor can appear anywhere in the nesting.

Conclusion

We addressed the question of the views that minds can have of themselves, rather than directly addressing the question of what minds are really like. However, we noticed that these views are part of the real nature of mind and may additionally be an intrinsic aspect of conscious qualia. Moreover, we conjectured

that entertaining a particular view of itself for a particular cognitive purpose can cause a mind to become more similar at around that time to how it is portrayed by the view. This could happen through at least two different mechanisms: roughly speaking, one aspect of the mind could distort another by acting on it in conformity with a metaphorical view, or one aspect could distort itself by coming more into conformity with a view of it employed by another aspect.

Although these points could be argued to apply to any sort of view, we concentrated on the special case of metaphorical views. Metaphorical views may be needed in practical self-reflection, just as they are needed in practical natural language discourse about mind, because of the messiness, complexity, and subtlety of mental states and processes. Metaphorical views throw into especially sharp relief the likely partiality and inaccuracy of individual views and the inconsistency between different views.

Our work on the ATT-Meta system for metaphorical reasoning provides ideas on how natural and artificial minds could think metaphorically, and thus make the general considerations of this chapter more real for mind researchers.

Acknowledgments

This research was supported by grant GR/M64208 from the Engineering and Physical Sciences Research Council of the United Kingdom. This chapter was developed from a talk created for Aaron Sloman's symposium *How to Design a Functioning Mind* at the 2000 Convention of the Society for Artificial Intelligence and Simulation of Behaviour (AISB-00), held at the University of Birmingham. The talk was given in adapted form at the EURESCO conference *Mind, Language and Metaphor: Euroconference on Consciousness and the Imagination*, Kerkrade, The Netherlands, in April 2002.

References

Baars, B. J. (1993). Why volition is a foundation problem for psychology. *Consciousness and Cognition, 2*(4), 281–309.

Barnden, J. A. (1997). Consciousness and common-sense metaphors of mind. In S. O'Nuallain, P. McKevitt, & E. Mac Aogain (Eds.), *Two sciences of mind: Readings in cognitive science and consciousness* (pp. 311–340). Amsterdam/Philadelphia: John Benjamins.

Barnden, J. A. (1998). Combining uncertain belief reasoning and uncertain metaphor-based reasoning. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society* (pp. 114–119). Mahwah, NJ: Lawrence Erlbaum Associates.

Barnden, J. A. (2001a). Uncertainty and conflict handling in the ATT-Meta context-based system for metaphorical reasoning. In V. Akman, P. Bouquet, R. Thomason, & R. A. Young (Eds.), *Proceedings of the Third International Conference on Modeling and Using Context* (pp. 15–29). Lecture Notes in Artificial Intelligence, Vol. 2116. Berlin: Springer.

Barnden, J. A. (2001b). *Application of the ATT-Meta metaphor-understanding approach to selected examples from Goatly*. Technical Report CSRP-01-01, School of Computer Science, The University of Birmingham, UK.

Barnden, J. A., & Lee, M. G. (2001). *Understanding open-ended usages of familiar conceptual metaphors: An approach and artificial intelligence system*. Technical Report CSRP-01-05, School of Computer Science, The University of Birmingham, UK.

Barnden, J. A., Glasbey, S. R., & Wallington, A. M. (to appear). Metaphor and truth from an artificial intelligence standpoint. In A. Burkhardt, & B. Nerlich (Eds.), *Reflections on tropical truth: Studies on the epistemology of metaphor.*

Barnden, J. A., Glasbey, S. R., Lee, M. G., & Wallington, A. M. (2004). Varieties and directions of inter-domain influence in metaphor. *Metaphor and Symbol, 19*(1), 1–30.

Barnden, J. A., Helmreich, S., Iverson, E., & Stein, G. C. (1994). An integrated implementation of simulative, uncertain and metaphorical reasoning about mental states. In J. Doyle, E. Sandewall, & P. Torasso (Eds.), *Principles of knowledge representation and reasoning: Proceedings of the Fourth International Conference* (pp. 27–38). San Mateo, CA: Morgan Kaufmann.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577–660.

Dennett, D. C. (1991). *Consciousness explained.* London: Penguin.

Fainsilber, L., & Ortony, A. (1987). Metaphorical uses of language in the expression of emotions. *Metaphor and Symbolic Activity, 2*(4), 239–250.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The Structure-Mapping Engine: algorithm and examples. *Artificial Intelligence, 41*(1), 1–63.

Gibbs, R. W., Jr. (1994). *Poetics of mind: Figurative thought, language and understanding.* Cambridge: Cambridge University Press.

Glasgow, J. I. (1993). The imagery debate revisited: A computational perspective. *Computational Intelligence, 9*(4), 309–333.

Grady, J. E. (1997). THEORIES ARE BUILDINGS revisited. *Cognitive Linguistics, 8*(4), 267–290.

Hobbs, J. R. (1990). *Literature and cognition.* Stanford University, CA: CSLI Press.

Horne, P. V. (1993). The nature of imagery. *Consciousness and Cognition, 2*(1), 58–82.

Johnson, F. (1985). The western concept of self. In A. J. Marsella, G. DeVos, & F. L. K. Hsu (Eds.), *Culture and self: Asian and Western perspectives* (pp. 91–138). London: Tavistock.

Johnson, M. (1987). *The body in the mind.* Chicago, IL: Chicago University Press.

Kövecses, Z. (2000). *Metaphor and emotion: Language, culture, and body in human feeling.* Cambridge: Cambridge University Press.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed.). Cambridge: Cambridge University Press.

Lakoff, G. (1996). Sorry, I'm Not Myself Today: The metaphor system for conceptualizing the self. In G. Fauconnier, & E. Sweetser (Eds.), *Space, worlds, and grammar* (pp. 91–123). Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Lee, M. G., & Barnden, J. A. (2001). Reasoning about mixed metaphors with an implemented AI system. *Metaphor and Symbol, 16*(1&2), 29–42.

Mio, J. S. (1997). Metaphor and politics. *Metaphor and Symbol, 12*(2), 113–133.

Mio, J. S., & Katz, A. N. (Eds.). (1996). *Metaphor: Implications and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Narayanan, S. (1997). *KARMA: Knowledge-based action representations for metaphor and aspect*. Ph.D. thesis, Computer Science Division, EECS Department, University of California, Berkeley, August 1997.

Rorty, R. (1980). *Philosophy and the mirror of nature.* Oxford: Blackwell; Princeton, NJ: Princeton University Press.

Stern, J. (2000). *Metaphor in context.* Cambridge, MA; London: Bradford Books, MIT Press.