PART II
TRACKING MINDS
et us turn now to the second key concept of this study: “metarepresentation.”1 Introduced in cognitive science in the 1980s, it has since gained wide currency among theory-of-mind psychologists and philosophers of mind and has recently become a subject of a wide-ranging collection of essays, *Metarepresentations: A Multidisciplinary Perspective*, edited by Dan Sperber. Sometimes described as “a representation of a representation,” a metarepresentation consists of two parts. The first part specifies a source of representation, for example, “I thought . . .,” or “Our teacher informed us. . ..” The second part provides the content of representation, for example, “. . . that it was going to rain,” or “. . . that plants photosynthesize.”

Or, to come back to our *Mrs. Dalloway* passage, the sentence describing Hugh’s pen—“It was still in perfect order; he had shown it to the makers; there was no reason, they said, why it should ever wear out; which was somehow to Hugh’s credit, and to the credit of the sentiments which his pen expressed (so Richard Dalloway felt) as Hugh began carefully writing capital letters with rings round them in the margin . . .”—is a metarepresentation with a specific source. That little tag, “so Richard Dalloway felt,” alerts us to that source, that is, the mind behind the sentiment. Knowing whose sentiment it is constitutes a crucial aspect of our understanding of the psychological dynamics of this particular scene and of the novel as a whole. Moreover, as I will demonstrate shortly, our tendency to keep track of sources of our representations—to metarepresent them—is a particular cognitive endowment closely related to our mind-reading ability.

---

1 WHOSE THOUGHT IS IT, ANYWAY?
This section’s discussion of metarepresentations draws on the work of Leda Cosmides and John Tooby, particularly their essay “Consider the Source: The Evolution of Adaptations for Decoupling and Metarepresentation,” published in Sperber’s collection. I will not try to summarize their carefully nuanced argument here; instead, I will adapt it and selectively quote from it for the purposes of explaining metarepresentation in fiction. To students of literature interested in learning more about our metarepresentational ability and its possible evolutionary history, I strongly recommend reading the original essay.

To grasp the importance of our capacity to form metarepresentations, let us imagine for a minute that we do not have this capacity, that is, that we can entertain representations, but we are not able to keep track of their sources. Let us consider how, thus circumscribed, we would conduct ourselves in the three following hypothetical situations, each of which involves our receiving a piece of information ranging from trivial, to fairly important, to absurd.

(1) Imagine yourself sitting in your office (which incidentally has no windows) and getting ready to teach a class. A colleague, named Eve, drops by and mentions casually that it is raining hard outside. You do not simply “save” this new information in your mind—the mind is not, after all, a computer; instead you assimilate it by integrating various inferences resulting from this representation with what you already know about the world, and hence modify your plans for future behavior. Or, to use Cosmides and Tooby’s terminology, the information about the rain is treated by your cognitive architecture as “architectural truth”; that is, it is “allowed to migrate . . . in an unrestricted . . . fashion throughout an architecture, interacting with any other data in the system with which it is capable of interacting.” Here are some examples of the thoughts arising in the process of such an integration/migration: “I’d better take an umbrella with me because it is a long walk from here to the building where I will be teaching”; “I should postpone making that announcement about the change in the syllabus until the second part of my lesson because many of my students will be struggling to find parking closer to campus and will be late for class”; “Peaches will be cheaper at the Farmer’s Market this weekend because the drought seems to be over, so I should stop by the bank tomorrow and take more cash with me when I go to the Market on Saturday”; etc. As we can see, the range of databases affected by the information provided by Eve is so broad as to be, in principle, infinite.

(2) Next imagine that during her short stay, Eve tells you that a recent addition to the department, named Adam, is a terrible person and a bad
colleague. She has known him from a previous job, and she remembers him as selfish, rude, and incompetent. Again, in the process of assimilating this new representation, you will let it affect all kinds of mental databases. For example, you may decide that you should try to avoid working on the same project with Adam and will in fact cancel the lunch appointment that you had with him for next week. You may further begin to think that your department must be really going downhill—look what kind of people they hire these days!—and so maybe it is time to start looking for another job.

(3) Finally, imagine that as Eve stops by, she informs you that it is raining golden coins outside. Once she leaves your office, you immediately call the department’s secretary to cancel your class. You can’t teach now: you have better things to do. In fact, the thought of an early retirement has just entered your mind; with all that gold falling into your lap, you may as well leave all the grading and committee work behind you. You frantically look around the office for suitable containers and, having found some, rush outside to gather as much of the golden rain into your bags as you can. Before you leave the office, however, you do manage to make a couple of other phone calls. For example, you contact a car dealer and tell him that you are ready to buy that Mercedes that you have seen on his lot, the one you always wanted but knew that you would never be able to afford. Now you can finally get in touch with that inner conspicuous consumer (unless, that is, the devaluation of gold ensues quickly).

(3a) It is also possible, however, that the information about the golden rain strikes you as so obviously absurd that you just ignore it. You do not take in that representation at all; you do not assimilate it with any of your knowledge stores; you nod politely as Eve tells you about it and simply forget it the moment she is out of your office.

But let us see how these situations change once we have our metarepresentational capacity back and thus are able to consider the source of any new information. The first scenario actually stays the same. If you have no reason to suspect that Eve is misleading you about the rain, you adjust your plans (i.e., about the umbrella, the classroom announcements, and the bank) accordingly. The second and third scenarios, however, are markedly different this time around. When you hear from Eve that Adam is a bad colleague, you feel understandably concerned, but you do not cancel your lunch with him and you do not start looking for another job. Instead you keep Eve’s information in mind but wait for further evidence that would either strengthen or weaken her claim. If several weeks or months later you find out that Eve has a long-standing grudge against Adam and that her
stories about him might well be untrue, and if, meanwhile, Adam has been impressing you as a perfectly amiable person and a good coworker, it is likely that you will revise that initial bad impression about him that Eve has saddled you with. At the same time, you will not just “discard” Eve’s communication as if it never happened; you will still retain the metarepresentation, “Eve told me that Adam is a bad colleague,” because now it tells you something important about Eve herself. (On the other hand, if some time later Adam does turn out to be a bad apple, you will come back to the information provided by Eve and consider it once more.)

Finally, in the case of the reported golden rain, once you have ascertained that your colleague is not being ironic (“yea, right, it will rain gold in front of a building housing the English department!”) or playing a practical joke on you, that is, once you are convinced that she is serious, you will take her representation, “it is raining golden coins,” in by integrating it with what you already know about the world. Only, this time, your inferences will focus mainly on this particular colleague and your future behavior in relation to her. You may decide to double-check any information that issues from her in the future, and you may consider not entering into any collaborative projects with her—just in case. Again, you will not just discard the incident (the 3a scenario). Discarding it completely could be dangerous because, however wrong, that information still tells you something important about Eve, something that you are better off knowing now rather than in the future when you are put in a situation in which you depend on her. Of course, in time, you may come to revise and abandon your suspicious attitude toward Eve and consider the “golden rain” remark a single instance of bad judgment or a silly joke; or you may come to believe, based on your later experiences with her, that she is indeed not very mentally stable.

In other words, our metarepresentational ability allows us to store certain information/representations “under advisement.” What it means is that we can still carry out inferences on information that we know is incorrect (e.g., “it is raining golden coins”) or have certain doubts about (e.g., “Adam is a bad colleague”) but that the scope of these inferences will be relatively limited. The “meta” part of the representation, that little “tag” that specifies the source of the information (e.g., “it was Eve who told me that . . .”) is what prevents the representation from circulating freely within our cognitive system and from being used as an input to “many inferential processes, whose outputs are inputs to others.” Instead of being available to all of our stores of knowledge and prompting us to
adjust our behavior in numerous ways, some of which could be harmful to us, that information is stored in what Cosmides and Tooby call a “suppositional” format and is thus available to a very selective set of cognitive databases, many of them having to do with the source of information. At the same time, “once [information] is established to a sufficient degree of certainty, source . . . tags are lost . . e.g., most people cannot remember who told them that apples are edible or that plants photosynthesize.”

The concept of metarepresentationality begins to figure in psychologists’ discussions of the difference between our episodic memories (i.e., memories tied to specific learning episodes or experiences) as compared with semantic memories (i.e., general knowledge not tied to specific learning experience). It has been suggested that “episodic memories are stored and retrieved via metarepresentations.” That is, such memories retain the time-, place-, or agent-specifying source tags and as such are stored as events that have been “experienced by the self at a particular and unique space in time . . ., with conscious awareness that ‘this happened to me.’” I may thus remember, for example, that it was last Thursday (the time-specifying tag), when I had dinner at my friend’s house (the place-specifying tag), that she told me (the agent-specifying tag) that I should try to use shorter sentences in my scholarly writing (the representation or memory itself).

By contrast, semantic memories are representations that are stored without the source tag:

Semantic memory . . . enables a person to have culturally shared knowledge, including word meanings and facts about the world, without having to recollect specific experiences on which that knowledge was based (e.g., knowing that Sacramento is the capital of California [or, to use the example above, that plants photosynthesize]).

Note, however, that a semantic memory—or a representation stored without any source tag—could acquire a source tag and become a metarepresentation. For example, people used to think that Earth was the center of the universe with other heavenly bodies orbiting around it. Gradually, however, this semantic memory, this culturewide, incontrovertible knowledge, became a metarepresentation with a source tag, “[P]eople used to think that. . . .” Moreover, we can append any semantic memory with a source tag and thus turn it into a metarepresentation, if only for the purposes of discussion, for example, “Lisa does not believe that Sacramento is
the capital of California.” By the same token, throughout our lives, we treat an untold number of semantic memories as absolute truths—for example, if you drop a shoe, it will fall—even though we can imagine conceptual frameworks within which these memories are not true anymore, say, in space, outside of Earth’s gravitational field. For practical reasons, however, it does not make sense for us to keep in mind all those alternative frameworks and thus store the representation, “if you drop a shoe, it will fall,” with a place tag such as, “on Earth” (unless we are astronauts). What these examples show is that although the distinction between the semantic and episodic memories (or between representations and metarepresentations) is useful both for our cognitive information management and for our discussions of cognition, this distinction is always context-dependent and potentially fluid.

Metarepresentational ability might have evolved in response to a very particular cognitive challenge faced by our ancestors. As Cosmides and Tooby point out, humans stand out “within the context of the extraordinary diversity of the living world” because of their ability to use “information based on relationships that [are] ‘true’ only temporarily, locally, and contingently rather than universally and stably.” On the one hand, this ability to make use of local, contingent facts fuels “the identification of an immensely more varied set of advantageous behaviors than other species employ, giving human life its distinctive complexity, variety, and relative success.” On the other hand:

The exploitation of this exploding universe of potentially representable information creates a vastly expanded risk of possible misapplications, in which information that may be usefully descriptive in a narrow area of conditions is false, misleading, or harmful outside of the scope of those conditions. Exactly because information that is only applicable temporarily or locally begins to be used, the success of this computational strategy depends on constantly monitoring and re-establishing the boundaries within which each representation remains useful. . . . Information only gives an advantage when it is relied on inside the envelope of conditions to which it is applicable.

As a constant monitoring and reestablishment of boundaries (e.g., “Adam is a bad colleague, but only in Eve’s representation of him”), our metarepresentational capacity is thus “essential to planning, interpreting communication, employing the information communication brings, evaluating others’ claims, mind-reading, pretense, detecting or perpetrating
deception, using inference to triangulate information about the past or hidden causal relations, and much else that makes the human mind so distinctive.” The lack of such an ability could be characterized as “naïve realism”—a state that Cosmides and Tooby suspect was “an ancestral cognition for all animal minds.” They further point out that although cognitive “systems of representational quarantine and error correction” that evolved to differentiate among representations, “storing” some of them with source tags that limit their scope of inferences, “are, no doubt, far from perfect [:]. . . without them, our form of mentality would not be possible.”

What is the relationship between Theory of Mind and metarepresentationality? Baron-Cohen and Sperber have argued, separately, that “the ability to form metarepresentations initially evolved to handle the problems of modeling other minds or the inferential tasks attendant to communication.” After all, consider how crucial it is for a social species such as ours to be able to attribute thoughts to people around us while keeping track of ourselves as sources of those attributions in case we need to revise them later (e.g., “I thought you wanted to go to the store with me because you got up from the table, but now I know that I was wrong: it seems that you just wanted to stretch a bit”).

On the one hand, Cosmides and Tooby agree with these views by stressing that the “restricted applications of inferences,” achieved by processing information metarepresentationally, is “not an oddity or byproduct of . . . [ToM], but . . . a core set of . . . adaptations essential to modeling minds of others accurately.” On the other hand, they point out that “the problems handled by metarepresentations . . . are so widespread, and participate in so many cognitive processes, that it is worth considering whether they were also shaped by selection to serve a broader array of functions—functions deeply and profoundly connected to what is novel about hominid evolution.”

Oliver Sacks’s research into the cognitive neuroscience of vision seems to support the latter view. Although Sacks does not use the term metarepresentation, here is a resonating quote from his essay describing the experience of an Australian psychologist, Zoltan Torey, who went blind at the age of twenty-one and since then worked hard “to maintain, if only in memory and imagination, a vivid and living visual world”:

Tory maintained a cautious and “scientific” attitude to his own visual imagery, taking pains to check the accuracy of his images by every means available. “I learned,” he writes, “to hold the image in a tentative way,
Because the argument of my essay focuses on literary texts, I have so far dealt and will continue dealing with verbal or verbalizable metarepresentations, such as, “Eve says it is raining outside.” Torey’s emphasis on holding *images* in a “tentative” way—visual metarepresentation, if you will—reminds us that, as Sacks points out, “there is increasing evidence from neuroscience for the extraordinary rich interconnectedness and interactions of the sensory areas of the brain, and the difficulty, therefore, of saying that anything is purely visual or purely auditory, or purely anything.” In other words, whether we agree with Baron-Cohen and Sperber, who think that metarepresentational ability evolved primarily to model human minds, or with Cosmides and Tooby, who suggest that its gradual emergence must have responded to a broader variety of cognitive challenges faced by our ancestors, it seems that its functioning today informs our interaction with the world on more levels than we are immediately aware of.

Part II: Tracking Minds

METAREPRESENTATIONAL ABILITY
AND SCHIZOPHRENIA

I have considered above three conjectural instances of our taking in *any* new information as an architectural truth. Now it is time to ask what really happens when the cognitive mechanisms that allow us to store information under advisement are damaged. A number of neurological deficits, such as autism and schizophrenia, have been linked to the failure of metarepresentational capacity, as have several kinds of amnesia. To begin with the mildest functional instance of such a failure, children develop a mature Theory of Mind around the age of four, and it is suggestive that just before that (typically, from three to four), they can go through so-called childhood amnesia, that is, a tendency to “believe that they actually experienced events that never happened, if they are asked about these *fictionitious* events repeatedly,” a consequence, perhaps, of having an immature “system for source tagging.” (This is not to say, of course, that as adults