Empiricism, Cognitive Science, and the Novel

Jonathan Kramnick
Rutgers University

> "I see into minds, you see," the robot continued, "and you have no idea how complicated they are. I can't begin to understand everything because my own mind has so little in common with them—but I try, and your novels help."
>
> —Isaac Asimov, *Liar!*

No one literary form has a proprietary stake in the mind, but as genres go the novel has since its inception taken remarkable interest in mental states. Among other things, eighteenth-century fiction is so much writing about the mind: about how thoughts represent things, cause other thoughts to happen, or lead to actions. The same might be said for empiricism. Seventeenth- and eighteenth-century philosophy paid unusual attention to the content of minds and the nature of ideas, to "human understanding" as Locke and Hume put it. While the connection between empiricism and the rise of the novel is a touchstone of literary studies, with a venerable tradition of scholarship dating back to the beginnings of the profession, only recently have critics drawn upon philosophy of mind and cognitive science to talk about the way in which thinking takes shape in particular works from the period.[1] This is of course not so much of a surprise, since criticism is as a rule skeptical of framing older texts with present-day models. The risk is one of anachronism or universalism, either shoehorning recalcitrant descriptions of the mind into our current language of cognition or locating both within a timeless and unchanging account of the psyche. Needless to say, my intention in this essay is to do neither; it is rather to consider what kind of insights can be gained by placing the description of thinking in the fiction and philosophy of the eighteenth century alongside certain tendencies within contemporary philosophy of mind and cognitive science—alongside, that is, the way in which we now talk about the mind. I'll begin with a comparison between empiricist and computational accounts of mental architecture and look at how each describes the shape and process of

cognition. I'll then turn to "theory of mind," a line of work in cognitive science that has proven especially attractive to literary studies because it concerns the way in which thinking about the thoughts of other people can be modeled or provoked by works of fiction.

1. MENTAL ARCHITECTURE: FROM THE ASSOCIATION OF IDEAS TO THE LANGUAGE OF THOUGHT

Despite their many differences, there is an important sense in which empiricism is compatible with cognitive science.[2] Most but not all philosophers of the seventeenth and eighteenth centuries had some sort of representational theory of mind; most but not all cognitivists do too.[3] On this view, the mind works by forming representations of objects and events and then implementing them in various processes of thought. "Concerning the thoughts of man," Hobbes writes in the first sentence of *Leviathan,* "they are every one a Representation or Appearance, of some quality or accident of a body without us; which is commonly called an object."[4] Slide ahead a few hundred years and things are not so different. "Mental processes," writes Jerry Fodor, one of the more influential and controversial philosophers of mind and cognitive science today, "are computations, that is, they are defined on the syntax of mental representations."[5] For Hobbes as for Fodor, the work of the mind is to have thoughts about or of some distal entity or state of affairs and then to put thoughts together in such a way that leads to behavior. Thoughts are "intentional" in the sense coined by Franz Brentano: one has a belief about one thing or wants another, and unless those things are other minds, the object of belief or desire does not have intentionality itself.[6] Hobbes found this point to be worth some emphasis; "the thing we see is in one place; the appearance, in another" and between the two lies some sort of reference or allusion (14). When "at some certain distance, the reall and very object seem invested with the fancy it begets in us," we ought to remember that "the object is one thing, the image or fancy is another," and we ought to recognize that images and fancies are matters of thinking, while objects and events are matters at which thoughts are directed (14). What our minds do is create images out of perceptions and memories, and once that is done they piece images together in the "succession of one thought to another" we experience as mental life (20). Fodor's argument that the mind is like a computer is thus at least in a preliminary sense compatible with the view presented by Hobbes. Each makes a case for intentional realism and claims that mental states have semantic and causal properties distinct from other states of the world.[7] Mental states are typically about something and in their aboutness have a peculiar capacity for meaning and for agency.

As we shall see, the architecture of mental representations differs sharply between empiricism and its computational descendants. Before taking up

these differences, however, I want to make two points in order to forestall potential worry. First, to say that a representational theory of mind distinguishes the mental from the nonmental is not to say that it holds a dualist or Cartesian view of the two. The distinction is one of function not one of substance. What the mind *does* is process information; what the mind *is*, ultimately, is an abstraction from matter.[8] Second, to speak of mental states in terms of intention and representation is not to say anything about what assumptions are attached to them by historical circumstance. Thus a representational theory of mind does not itself entail any loose talk of subjectivity, privacy, interiority, selfhood, the individual, autonomy, the human, or even (save for Locke) consciousness.[9] Any one of these things may (or may not) be joined with whatever account of thinking one prefers, but there is nothing in the logic of either empiricist or computational accounts of mind that requires a particular version of the entity in which thoughts are ostensibly occurring. And with these caveats, let us turn to the historical matter at hand.

One distinguishing feature of empiricism was that it attempted to cobble together an account of thinking with an account of epistemology. For Hobbes as for Hume, the architecture of mental representation was meant to satisfy a set of concerns about knowledge. When we think, our ideas tell us something about an external world. That world exists independently of our thoughts yet can only be understood through the images we have of it in the mind.[10] Consider Locke's famous description of the mind as a kind of camera obscura: "the Understanding is not much unlike a Closet wholly shut from light, with only some little openings left, to let in external visible Resemblances, or *Ideas* of things without; would the Pictures coming into such a dark Room but stay there, and lie so orderly as to be found upon occasion, it would very much resemble the Understanding of a Man."[11] On this description, the mind is a three-dimensional place littered with images, and thinking is a "repeating and joining together" of images into a sequence or *association* in which one idea leads to the next by means of some sort of inference (2.12.17). To have a mental state is to be in view of a representation, a picture of something one experiences or a series of pictures one puts together, and to be in view of a representation is to be in some relation of greater or lesser accuracy to a world that is being depicted. Locke's project, in this last respect, is not so much to understand what our beliefs are or how they are structured as it is to figure out whether we are justified in having the beliefs that we do.

In contrast to empiricism, the computational model tends to split apart such questions of epistemology from questions of psychology, to be less concerned with the accuracy of representations than with how they are put together and thus to consider ideas as something like units of a mental language.[12] Each has a theory of mental representation, but the nature of cognition differs between them. If the one associates over a parallel sequence of ideas, the other performs operations over a structured order of symbols. The

mind is not so much a screening room for pictures, on the latter view, as it is an instrument for processing concepts. Consider the following thought: "Jonathan is writing an essay." To think that Jonathan is writing an essay, according to the computational logic, isn't to think Jonathan and then writing and then essay. It is to think *about* Jonathan *that* he is writing *an* essay. The predicate "is writing" holds the entities "Jonathan" and "essay" in such a way that some parts of the thought have priority over others in determining its overall meaning. For the thought to be a thought, on this view, it needs to be more than a series of pictures, or else we would never be able to think Jonathan "is writing" rather than the meaningless series "Jonathan/is/writing . . ." or the meaningless lumps "Jonathan is," "writing an," and so on.[13] Like a computer, the mind is sensitive to the syntactical arrangement of its concepts. Unlike a computer, however, the mind is not sensitive only to syntax. For the semantics of a thought to arise from its configuration, each of its lexical units must also have an independent meaning. On the computational model, therefore, thinking tends to have what linguists call a *compositional* form, according to which the meaning of any one thought is determined by the structure of its syntax and the meaning of its constituents.[14]

The point of the comparison is not so much to scold empiricism as an inadequate philosophy of mind as it is to indicate what is important in its cognitive architecture by means of its difference from current models of mental representation. I will argue later in this essay that what is allegedly wrong about empiricism turns out to be "right" for the novel, although not perhaps in the way that we have traditionally imagined. To get us there, let us observe the way in which Fodor describes the difference: "mental representations are sentence-like rather than picturelike. . . . In sentences, there's a distinction between *mere* parts and constituents, of which the latter are the *semantically interpretable* parts. By contrast, every part of a picture has an interpretation: it shows part of what the picture shows."[15] Fodor's point is that empiricism's commitment to the image entails that it can only consider thinking as one idea after another and not, as he would prefer, a computation over the whole. "Associations are operations on parts of mental representations," while "computations are operations defined on their constituent structures."[16] The distinction has an important corollary for the way in which we consider the mind. An association can never get beyond the image because it ties semantics in an epistemic relation to objects. In contrast, computations discover the semantics of mental representations in the way in which their lexical units are organized.

Across the long divide between empiricist and computational theories of the mind, therefore, several important distinctions come to the fore. The computational model agrees with the empiricist model that ideas exist in our mind as representations, but disagrees with the empiricist corollary that representations are pictures of things.[17] Because of its epistemic commitments, empiri-

cism wants to argue that simple ideas are copies of experience and that complex ideas are built from relations among simple ones. Hume, for example, proposed not only that impressions lead to ideas but also that simple ideas lead to complex ideas in virtue of the regularity of connections between them. "Were ideas entirely loose and unconnected," he writes in the *Treatise,* "chance alone would join them; and it is impossible the same simple ideas should fall regularly into complex ones (as they Commonly do) without some bond of union among them, some associating quality, by which one idea naturally introduces another."[18] This quality requires some sort of inference from one idea to the next, either resemblance, contiguity, or cause and effect, but in every case the relation is connective and probable with nothing that would add to or break apart the idea-images themselves.[19] It is this connection of fused parts that is so difficult to bring into computational theory, according to which (again) the meaning of a complex idea inheres in its constituent structure not in its order of pictures. "If images are to serve as vehicles of thought," writes the cognitive psychologist Zenon Pylyshyn in a recent study of vision, "they must have what might be called interchangeable parts, much as lexical items in a calculus do."[20] And if images are to have interchangeable parts, they aren't exactly images as we are accustomed to thinking of them; they would be "more language like than pictorial" and would lose their "alleged depictive nature."[21] So while we may think we think in pictures, in fact we think in something closer to a script, one that has the formal capacity to encode representations in a computable syntax.

It is with some surprise, then, that one reads in Fodor's recent monograph on Hume how "Hume's *Treatise* is the foundational document of cognitive science."[22] The effort is to revive Hume by purifying his philosophy of its empiricism. Fodor is favorably disposed to the *Treatise* because "it made explicit for the first time the project of constructing an empirical psychology on the basis of a representational theory of mind."[23] The only trouble with Hume is that his psychology has refused to let go of his epistemology. There is no reason to suppose that concepts are tethered to the impressions that provide their warrant. The impression for example of a dog could plausibly decompose into the concepts of dog or animal or mammal or quadruped or Snoopy. One distal object has an array of possible simple concepts, all of which could then become the sentential units of complex concepts. So Hume is right to say that complex thoughts are built from simple and irreducible ones, just wrong about the copying from impressions and about the manner of construction.[24] The associative structure carries much of the blame. Hume wants complex ideas to contribute to the meaning of their constituents: I have never heard Snoopy bark but I can formulate the idea of Snoopy barking by connecting my image of the first to my image of the second. But how then is this anything other than placing two simple ideas in a sequence, and how is thinking anything more than hanging the same pictures in different places? Computation wants to

suggest that simple concepts are, as it were, simpler than impressions and complex concepts more complex than associations. Without both elements of the theory in place, we cannot explain how minds go about thinking: "Hume needs an argument that the structure of complex concepts is semantically transparent, so that if the content of the simple constituents is experiential, then so too is the content of complex concepts constructed from them. But he clearly hasn't got such an argument, and since the semantic productivity of novel concepts requires their structure not to be semantically transparent, I can't imagine where he might look for one."[25] We are thus, on Fodor's account, witness to an interesting failure. The association of ideas is transparent because its merely causal structure doesn't contribute anything on its own: Snoopy merely runs into or precedes barking. Yet structure is precisely what is needed to create new thoughts (Snoopy barking, as the case may be). "There is a tension," Fodor writes, "between what semantic productivity requires and what empiricism permits; the former wants the structure of a representation to 'add something' to the content of its constituents, but the latter wants it not to. Well, since productivity isn't negotiable, maybe Hume should give up on his empiricism. Come to think of it, maybe he should give up on trying to infer his epistemology from his psychology. Come to think of it, maybe we should all do that."[26] Fodor's elegant drollery should not obscure the salient difference in view. The difficulty with Hume's empiricism is that it continues to derive complicated ideas from associations of simple ideas and to define simple ideas as copies of experience. Despite their ostensible simplicity, these idea-images are just too big and inflexible to fit into a sentence of a mental language, and so therefore Hume has no way to cash out the cognitive science he invents. "Bother epistemology," Fodor concludes. "And bother empiricist epistemology most of all."[27]

Those of us who are less interested in whether empiricism accurately described the way in which the mind works than in what its model of the mind can tell us about seventeenth- and eighteenth-century culture might still learn a great deal from the frustration in this particular area. If the mental token in the representational architecture of empiricism turns out on comparison to be an image of an object and thus to be object*like,* what for example might this tell about the related model of agency, or as we now say, of mental causation? The computational model tends to consider behavior as an output of a particular type of propositional attitude, an operation that takes the sentential form of "desire that . . ."; on this account, to say that I desire some particular object is to say that the object of desire is not properly an object at all (or at least not an image) but rather a constitutive unit of a language whose semantic yield supervenes on its syntactic placement. Empiricism also talked about agency in terms of attitudes, but with an important difference. There the representational structure preserved the pictures of objects without decomposing them into their constituents.[28] Locke, for example, argued that what

motivates a person to take this or that action is uneasiness in the want of some absent thing. One acts to relieve this uneasiness by striving toward an object of desire or by ridding oneself of an object of distaste.[29] The arc of the mental state that produces action is propositional (or aspires to be), but the structure in which the proposition is expressed is not a language. One's attitude is in relation to an object-image, not to the sentence in which that object-image is decomposed. And so agency on this view is an output of a person's relation to a mental token that is picturelike. In taking account of a person's actions, we ought to infer backward to some prior mental state in which that person stood in relation to an internal object in roughly the same manner that one stands before a thing one sees.

When writers like Locke and Hume attempted to account for knowledge, they thus had to account for what it *felt* like to stand before an image. On the computational side of things, the question of knowledge ought really to be dispensed with when we talk about thinking. It can only introduce a fateful confusion of epistemology with psychology, a discussion of justification with a discussion of mental process, and thus a thralldom to the image instead of a manipulation of symbols. (It also introduces the messy question of who or what is experiencing the feeling of thought, a problem which computation tends to consider unanswerable and beside the point.)[30] In this instance, however, the confusion leads to an interesting dilation on the subjective experience of objects and associations. The cognitive disciplines would refer to this experience as qualia—the immediate sensory impact of an object—or when strung together as phenomenal consciousness.[31] Locke was indeed one of the first philosophers to name the relation we have to representations as "consciousness," and it is on this basis he argued in his chapter on personal identity that we have an idea of ourselves as "selves."[32] Locke's chapter on personal identity was one of the most controversial sections of the *Essay*.[33] It is also quite a familiar one to scholars of eighteenth-century literature, so I won't belabor its contents here except to remind us that Locke makes the point that personal identity resides in the awareness a mind has of its internal repertoire of experience: "consciousness always accompanies thinking, and 'tis that, that makes every one to be, what he calls *self;* and thereby distinguishes himself from all other thinking things, in this alone consists *personal Identity*" (2.27.9). To be a person is to have a series of connected experiences during which time one was aware of the representational nature of one's thoughts; it is to believe that the same person was in view of representations in the past as in the present, and that one ought to care about the person's fate in the future.

One of Locke's earliest readers, the philosopher, playwright, and novelist Catherine Trotter, takes up this point in her 1702 *Defence of Mr. Locke's Essay on the Human Understanding*. The argument Trotter feels the need to defend at length in the pamphlet is Locke's notion that identity resides in a form of consciousness defined as the awareness one has of representations. "*Personal Iden -*

tity," Trotter argues "consist[s] in the *same Consciousness* and not in the same *Substance,* for whatever Substance there is without *Consciousness* there is no *Person*."[34] So far Trotter adds little to the language of the *Essay.* The distinctiveness of Trotter's approach becomes apparent when she moves to describe how the self takes shape as a series of attitudes taken in relation to objects and thus, she reasons, as a series of objects itself. The mind is a peculiar kind of stockroom. "I am thinking," Trotter writes, "of a horse; his beauty strength and usefulness. Does this thought preserve the Idea of a Church, of Happiness or Misery," or for that matter of an "apple" or a "table"? Or are these things pushed out by new ideas? "If they remain in the mind when I was only thinking of a horse," Trotter continues, "wherever they are bestowed, it may be presum'd, there is room for that one idea more without thrusting out another to give it place" (33–34). Because I still have the idea of a church somewhere in my mind when I think about a horse, or because I can preserve the idea of a table or a person when I think of an apple, I am able to string different objects into unique thoughts. And because I am able to produce and be conscious of such thoughts, I am the same person today as I am tomorrow, one subject to the distinctive accidents of life narrative and culpable for my actions in this world and the next.

The leap to identity is, as Locke's critics pointed out, not so much proven as assumed in this argument, since there is nothing in the logic of mental representations to entail that the same person is always in view of different ideas.[35] Trotter takes it upon herself to make ideas supervene on a person who is having them. Her consideration of whether there is space in her head for more pictures—whether she can find room for an apple once she looks at a horse—is in this respect a kind of literal version of the theory of association, a way of imagining one's life as a constant series of mental representations. Ideas are tokens of things and also things in one's mind. The mind is a bottomless satchel of ideas. Life is an aleatory string of connections between them. A writer of fiction as well as philosophy, Trotter was well-suited to draw out the supervenience of ideas upon persons. Around the same time as Locke's *Essay,* she published the epistolary novel *The Adventures of a Young Lady,* a loosely told story of the various amours of Olinda who one day meets and falls in love with the older Cloridon. Deceived into believing that Cloridon has forsaken her for another, Olinda finds herself rather more troubled than she would have expected. "I found myself seiz'd with an unusual I knew not what."[36]

> As soon as I was alone, I examin'd my self upon the matter. Why shou'd this trouble me (said I within myself) who wou'd not entertain his Love, when it was offer'd me, and I have often Resolv'd never to see him, even when I thought him Constant? How comes it then, that I am so Griev'd and Angry that he loves another? And that I wish with such impatience for his Return? In fine, I discover'd that what I had call'd Esteem and Gratitude was Love; and I

> was as much asham'd of the Discovery, as if it had been known to all the World. I fancy'd every one that saw me, Read it in my Eyes: And I hated my self, when Jealousies would give me leave to Reason, for my extravagant thoughts and wishes. (66)

Olinda's bout of concern leads her to inspect the procession of her ideas and so arrive at what Trotter and Locke would call her personal identity. The "I knew not what" that seized her can only be revealed by introspection and self-reporting. Thus epistolary disclosure is augmented by internal speech, as if the writing down of ideas was not enough to display their actual content and she needed to strip the external covering to see the mental process itself. The voice inside Olinda's head represents her ostensibly real thoughts, the voice of the letter writer a reflection on those thoughts that leads to a discovery of their truth (that she is in love). Olinda's ideas become so clear to her as pictures that she imagines others must be able to see them as well. Or so she fears when self-examination leads her, at the end, to place herself in the position of someone else viewing the young lady named Olinda who is so clearly in love with Cloridon. The multiple perspective slows down the train of Olinda's ideas so that each may be separately examined as a discrete image; this is how she discovers that esteem and gratitude are really love. As Trotter will go on to argue in defense of Locke, the self is a collocation of mental states, a collocation to which we may attribute desire, belief, and finally, action. Understanding that a person's identity is composed of thoughts that have the property of images and images that are tokens of objects thus credits that person with an inner life and accounts for her behavior.

2. MIND READING

Olinda's letter is an object in a series of objects, and so it both peers into and provides a model for her thoughts. The unusual proximity of Locke and Trotter, and the interesting anomaly that Trotter was a writer of both philosophy and fiction, reveal a more general point about empiricism and epistolary form at the turn of the century: each is interested in putting the interior states it reveals into a sequence. Epistolary novels place one letter after another, and like the minds they represent run by association not by computation. (As a model of the mind, *Pamela* has no language of thought.) They also draw attention to one special feature of the variety of objectification on view in the passage from Trotter's story, namely, that the particular object represented within the mind of Olinda is her own mental state, by definition not something observable in the external world. Since all mental states are representations, the object-image we are invited to examine is a metarepresentation of what she is thinking, a token of a token. Such high-order metarepresentation is relatively common for fictional characters, and can occur in moments of both

introspection and self-reporting—Olinda's thoughts about what she must be thinking—and in reflection on the minds of other people. Anytime one character attributes ideas to another or thinks about what another character might be thinking, the mental process involves turning that person's thoughts into an object within one's own mind. Contemporary philosophy and cognitive science calls this process "theory of mind," so named because it describes how one mind goes about developing a theory of the contents of another (or, as in the case of first-person mind reading like Olinda's, of itself).[37] The point is that mental states are never observable in the same manner as Trotter's horse, and so some sort of inferential process, or mind reading, must be at work in order for them to be represented or formed into tokens at all. This is especially so in the case of third-person mental states where attribution is undertaken at a distance from the mind that is being represented. The eeriness and power of the Olinda passage, for example, derives in part from her reading of her own mind from a third-person perspective, imagining that some outside observer could view its contents through the look in her eyes.

The sort of theory that Olinda is developing in this passage is of her own mind and it is presented to us as a discovery that she is in love. The manner in which the passage slows down to present the process of mind reading and then switches at the end to a third-person perspective on first-person experience describes a rather intricate version of what is elsewhere an ordinary procedure, one in which, as Alvin Goldman puts it, "we attribute a host of mental states to self and others" and interpret human actions in the "mentalizing" terms of desire or belief or intention.[38] The narration of Olinda's discovery seems so remarkable, in fact, because it is such a complicated and involuted version of an everyday practice. Her worry at the end about the visibility of her mental states is a doubly embedded act of representation, in which she reads what she fears other people are reading about her thoughts. To put this in terms of Trotter's empiricism, Olinda tokens her past mental state when she is alone and reflecting on what had seized her. This metarepresentation includes as it goes along an embedded mental state that belongs to "everyone" who, in turn, token the mental state of Olinda in love. Mental state representation is not always so recursive or multiply embedded, and in fact theory of mind as a philosophical or cognitive enterprise is designed to address the often simple ways in which agents interpret others or themselves in terms of belief or desire or one or another emotion. So for example when Olinda spurns an earlier suitor Berontus, she describes how he "left me almost as Angry at himself as he was before at me; and did not come near me for some time thereafter" (19). Olinda here attributes an emotion to Berontus because of the way he speaks and acts in her presence; she then draws inferential conclusions about his subsequent behavior (his not visiting her) on this basis even in the absence of any observable evidence. Fodor would call this "a piece of implicit, non-demonstrative, theoretical inference," not least because the episode is so

ordinary and unreflective.[39] One imagines that Trotter did not give the description a moment's pause, which is only to say that even the least remarkable incidents of third-person mind reading involve an objectification of another person's ideas so they may be formed into a token within the representational architecture of one's own mind.

I don't think that Trotter's novel is remarkable in this particular case, nor do I think it really could be. What I want to suggest, though, is that there is a relation between the empiricist model of thought and the way in which theory of mind problems were developed in the early novel. We have already seen that empiricism tends to model cognition in terms of association and semantics in terms of object-pictures: a horse giving way to an apple or a church or a concept of faith or of happiness or misery. There was something particularly amenable about this representational theory of the mind to the type of metarepresentation that goes on when one character attributes thoughts to another or when a reader attributes thoughts to both. One person represents another person's representation, and in each case the token is experienced as a kind of object-picture or an association among them. "He left angry" is an instance of second-order attribution and so a relatively simple instance of mind reading. Things get more complicated when one is attributing a mental state to someone who is doing the same. Suppose Olinda notices that Berontus is angry about her loving someone else: her mental state includes a token of his, yet his includes a token of hers. In recent work on theory of mind problems in the novel, Blakey Vermeule and Lisa Zunshine have each described the way in which such embedded "orders of intentionality," as Vermeule puts it, increase "the cognitive load on both writer and reader alike" in such a way that leads to experiments in literary form designed to capture the "gauzy filaments" of consciousness in the "fragile casing of narrative."[40] I will have more to say about this "cognitive load" below, but I would point first to the way in which the broader insight finds support in Trotter's representative version of epistolary thinking, with its sequential ordering of ideas and concern for first- and third-person mental states.

We need not look only at epistolary fiction or concentrate on the canonically "psychological" novels of Richardson or Burney in order to find intricate, multiply embedded orders of intentionality in the fiction of the period. Consider for example the following passage from Defoe's *Roxana,* where the eponymous heroine discusses the possibility of regret with her lover at the time, the French Prince: "My dear, says he, if once we come to talk of Repentance, we must talk of parting." The exchange then becomes speechless, contained within a single recursive thought: "If tears were in my Eyes before, they flow'd too fast now to be restrain'd, and I gave him but too much Satisfaction by my Looks, that I had yet no Reflections upon my Mind strong enough, to go that Length, and that I could no more think of Parting, than he could."[41] The burst of emotion suggests a transparent feeling, as if her disinclination *were*

the actual tears. Yet the content of her thought, as Defoe narrates it, is an intricately layered series of embeddings, reaching at the end to fourth- (and, if one counts the reader, fifth-) order intentionality. What Roxana writes might be reworded like this: "I realized that he thought that I had no more intention of leaving him than I thought he did of leaving me." Or even more boiled down: "I thought that he thought that I thought the same thing about him that he thought about me." Putting it the first way flattens out the passage, and putting it the second denatures it entirely. Even so, the two rewordings allow us to see the way in which Defoe attempts to compress within a single image a multilayered embedding of one mind within another described to a third.

To illustrate the passage from Roxana within the empiricist architecture of mental representation all we would have to do is make a slight adjustment and say that Roxana represents the Prince's token of her, which is itself embedded with a token of him, and that each token runs on an association of object-pictures, including especially the final moment when the tear contains both Roxana's feeling and her sense of what the Prince must be feeling. To what degree is it unhistorical to describe this as a theory of mind problem and to make recourse to cognitive science for its explanation? Not so much, I think, or at least not yet. Mental content (one's own and others') was an intense concern for the period that developed both the representational theory of mind and the literary genre in which the theory is most fully explored. Theory of mind, as I've described it so far, works as well as it does with the architecture of mental representation, in other words, in part because each is an eighteenth-century preoccupation. Roxana's tear is an output of internal relation to a mental token. That token has the quality of a picture of the Prince's feelings looking at a picture of hers. When we interpret her burst of tears, we attribute to her a mental state she describes as a feeling brought about by a certain image. What Fodor might describe as the error of Trotter's or Defoe's empiricism, then, might also be viewed as the way in which writers from the period formalized problems of thinking, mind, and mind reading.

Considering how entwined theory of mind is with the literature and philosophy of the period, I don't think it's very surprising that the cognitive disciplines have become of interest especially for eighteenth-century scholars, like Vermeule and Zunshine. Empiricism's attention to the cognitive solicits a notice from critics who then use theoretical tools in historical continuity with the theory of the period itself. In posing this as a historical problem, however, I am reading somewhat against the grain of their work. Zunshine's recent book, *Why We Read Fiction: Theory of Mind and the Novel* (2006), argues that novels are "grist for the mill of our mind-reading capacities" because they ask us to "posit a mind whenever we observe behavior as they experiment with the amount and kind of interpretation of the characters' mental states that they themselves supply and that they expect us to supply."[42] Our mind-reading capacities are, on this view, the same as any reader's in the eighteenth century.

When, for example, Zunshine argues that *Clarissa* is a "massive and unprecedented in Western-literary-history experiment with readers' Theory of Mind," she means to suggest that the novel's tragically entangled and often mistaken spots of attribution are unprecedented while the cognitive capacities of readers are not.[43] Thus the novel "reenters culture with every new interpretation because it is peculiarly geared to its exclusive environment," a fixed ecology that it "latches onto."[44] Fictional minds are one thing and real minds are another, and the one develops techniques to approximate or provoke the other.[45]

In place of this kind of argument, I've attempted to show how some of the formal features of the early novel match up with the naïve theory of mind seen at work in the fiction and philosophy of the period. Putting matters this way, though, risks contrasting a weak historicism to a hard-nosed cognitive science. We wouldn't want that; so let's look a little closer at Zunshine's claims. If the account of *Clarissa* seems to place the novel within the ostensibly permanent features of human psychology, that may be because of the particular strain of cognitive science with which she is working. Heavily indebted to the work of Simon Baron-Cohen, Zunshine has taken on board a series of assumptions about the evolutionary history and modular structure of the mind/brain over which there is considerable debate within the theory of mind literature itself.[46] Mind reading on her account is an adaptive capacity that "must [have] developed during the 'massive neurocognitive evolution' that took place during the Pleistocene (1.8 million to 10,000 years ago)."[47] As early humans began to live in groups, pressure was placed on interpreting behavior in terms of mental states, and so natural selection favored a cognitive architecture structured for metarepresentation. Theory of mind subserved a kind of atavistic chess, according to which hominids were always trying to figure out, in Baron-Cohen's words, whether someone's "next action is to attack you, to share its food with you, or to mate with you."[48] The logic of this argument works by what is often called reverse engineering, taking the alleged properties of the human mind and tracing them backward to some early moment in the Darwinian drama.[49] So according to this story, adaptation promoted the development of a Theory of Mind Mechanism (or ToMM), located in a special module devoted to inferring, or theorizing, mental states.[50]

The evolutionary-modular account is often described as a "theory-theory" because it emphasizes the way in which agents make speculative inferences about mental states.[51] Once the ToMM goes online, it supplies a rudimentary theory of mental content and allows agents to read behavior in terms of belief, desire, and the like. Our minds are designed, in this view, to provide automatic attributions according to one or another psychological law. ToMM receives as input certain information about a target's eye movement or speech or physical action and provides as output a token of what the mental state of the target is likely to be, in roughly the same way that a cyclotron measures particle speed. Mental states are themselves unobservable, yet we are built with mech-

anisms to provide inferential or theoretical knowledge of them anyway. When Zunshine writes of "the relationship between our evolved cognitive capacity for mind-reading and our interest in fictional narratives," therefore, she suggests that novels raise to the level of conscious apprehension a process soft-wired into a specific, encapsulated domain of the cognitive mind, one put in place by natural selection thousands of years before the writing of novels themselves.[52]

Zunshine's point is not exactly to reveal the novel as a fortunate holdover from a template set in stone ages ago. She wants to claim that there are an infinite number of possibilities that could arise from our cognitive endowment and that any particular genre must be traced to the circumstances of its time before it is correlated with the adaptive structures of the mind.[53] Even so, the burden of transposing this account of evolutionary development to the artifacts of a given historical moment—not simply the eighteenth century, but *any* spot of cultural time after the invention of writing—is considerable. It is a long way from the Serengeti Plane to Harlowe Place, and one wants to keep the "fragile casings" of narrative from buckling during the ride. Zunshine attempts to resolve this problem by emphasizing that theory of mind is "context dependent."[54] The module only comes online through social interactions, including the various tasks we perform as readers; while we do not need to read *Clarissa* to learn how to attribute mental states to other people or ourselves, our capacity to do so is given a workout when we do. Suggestive as this argument is, it still leaves the process of attribution entirely separate from the artifacts with which it is correlated. The one is literally prehistoric, the other from some moment in recorded time. Seen this way, the particular cast of any one novel supervenes upon a relatively inflexible structure of mind reading. After all, the three-hundred-or-so-year history of the novel measures not a microsecond in evolutionary time.

Where does this leave the episodes of mind reading we observed in Trotter and Defoe, let alone the drawn-out inferential drama of *Clarissa*? When Roxana attributes to the Prince a mental state that contains an image of her own embedded with his, she seems to interpret his behavior in light of her own feelings. He is every bit in love as she, which she knows because she feels so in love herself. The word for this sort of mental transposition during the eighteenth century, whether of propositional attitudes or emotions, is sympathy. I raise the connection here because eighteenth-century models of sympathy bear something of a family resemblance to the main rival of the theory-theory approach to mind reading, the "simulation theory" developed by philosophers like Goldman and scientists like Vittorio Gallese. Simulation theory, according to Goldman, "says that ordinary people fix their targets' mental states by trying to replicate or emulate them. It says that mindreading includes a crucial place for putting oneself in others' shoes."[55] The important point of distinction between this account and the theory-theory of Baron-Cohen and

others is that simulation presumes that agents only come to a theoretical inference of the contents of other minds after they first take their own system off-line and run a simulation routine of the target's mental state. Agents operate "their mechanism on the pretend input appropriate to the target's initial position [and] use their own minds to 'mirror' or 'mimic' the minds of others."[56] Inferential metarepresentation is thus the final output of a process that includes an initial, introspective self-token generated by an enactment of the conditions under which the other mind is understood to be thinking.

Third-person attribution on this account begins with a first-person simulation of the thoughts of someone else or oneself. While philosophers like Goldman are more reserved in their language of modules and less committed as a rule to evolutionary psychology, they share theory-theory's contention that mind reading typically occurs at the "functional or neural" level and only on occasion reaches the threshold of conscious awareness.[57] Were we to correlate the passage in *Roxana* to a simulation routine we would thus have to account for the way in which the novel slows down, in order to make explicit and track, a subpersonal and speedy cognitive mechanism.[58]

Perhaps that is what literature does after all, perhaps not. I would merely note here that the eighteenth-century's version of mind reading does not seem entirely apart from the version offered by simulation theory. Compare Goldman's account of shoe wearing to the opening page of Adam Smith's *Theory of Moral Sentiments* (1759). "As we have no immediate experience of what other men feel," Smith begins, "we can form no idea of the manner in which they are affected, but by conceiving what we ourselves should feel in the like situation."[59] Our senses can never "carry us beyond our person," so in order to form an image of what another person is thinking or feeling we must use our imagination to "place ourselves in his situation" and "conceive ourselves enduring all the same torments" as him (9). Smith's opening move is strangely familiar: mental states are by their nature perceivable only through some sort of inferential stance. Even if our brother is on the rack, we come by our sense of his thoughts indirectly, and even then it will only be a simulation generated by taking our system off-line and replicating his predicament in our mind. The resemblance of the simulation routine to the way in which our period talked about the mind contains no easy lesson, however, for how we might correlate the literary historical materials to the methods of cognitive science. One might want to say that it provides a way to avoid the transmillennial gap between cognition and culture we encountered in evolutionary theory-theory. Seen this way, mind reading need not arrive so hardened and so entirely in advance of fiction. The roots of the routine in eighteenth-century philosophy are, like the novel, part of an attempt to come up with a model of the mind. Yet, this risks collapsing the two into a facile symmetry, in which Goldman is preferable simply because he seems *like* the artifacts we are used to reading. Whatever we gain by identifying a shared project between our period of study and present-

day theory ought not to be purchased at the expense of what we have seen as the productive friction between the two or stand in place of doing the hard work of relating what happens visibly in novels and what might occur intractably in minds. In what remains, I'll sketch out some preliminary thoughts along these lines.

Those of us who work in literary study are, needless to say, not in a position to judge the relative merits of simulation theory and theory-theory as accounts of the mind itself. We ply our trade at the level of heuristics, putting modes of analysis together and seeing what emerges in the process. In this way, the appeal as well as the disadvantage of the simulation approach consist in the relation it illustrates between fictional and the philosophical versions of the mind and between the eighteenth-century materials and contemporary theory. The questions raised by this relation are, accordingly, under what sort of social, technological, and cultural pressures did the period come up with a model in which introspective mind reading became both possible and urgent, and according to what formal devices did writers evoke and render palpable a process understood to be mental and imperceptible? I cautioned earlier against tracing an evolution of literary forms in terms of a steadily more accurate account of what is set in stone in the distant past. Smith's version of the simulation routine provides, in this respect, an alternate model of how the language of cognitive science might cash out for literary study. When he writes that we have no way of getting beyond our own sense of things—that the first-person experience of our brother on the rack is inaccessible to us by anything other than an inference—he places great importance on the role of the imagination to reveal a second-order representation of what our brother might be thinking. We have seen already that what then ensues is a simulation that outputs a sympathetic sense of his pain.[60] The function of the imagination in its root sense to present images remains of interest. Smith here relies upon the empiricist account of cognition as an association of ideas rendered as object-pictures of experience: "It is the impressions of our own senses only, not those of his, which our imaginations copy. By the imagination we place ourselves in his situation . . . and thence form some idea of his sensations" (9–10). The imagination forms a picture out of one's own repertoire of experience in order to form a picture out of someone else's. We can imagine someone else's sensation if we manage to first imagine our own in her place. Arranged this way, the imagination allows for a kind of counterfeit sense, or picture taking, of one's experience as a replicated state of another's. In so doing, it relies on the implicit notion that mental images have the ability to yield intentional content, one's own or others.

The reliance on this sort of semantic yield is simply one error of empiricism, according to the computational model. The imagination is an insufficient medium for mind reading because the pictures it furnishes do not mean anything until they are decomposed into a language of thought. Or, to put it

another way, Smith places too much faith in the unbroken integrity of images, since after all, only some parts of an image are constituents of meaning and those are important in virtue of their syntactic placement. I raise the contrast of the associative structure of sympathy with the computational architecture of the mind at this point because it uncovers several clues about the complicated role the imagination plays.[61] As we have seen, the primary function of the imagination appears to be to provide an image of another person's thought or feelings. One can never experience what is in another person's mind, but one can have a second-order inference of it by imagining oneself in that person's place. In this respect, the imagination adds something to the ordinary association of ideas. The association pattern typically runs by putting together ideas that are copies of experience. Yet, mind reading involves producing an idea that has no correspondent in experience beyond what we can imagine were we to be in the same situation ourselves. The output copies a fictional token idea. At the same time that the imagination adds something to association, however, it also implements association. "When two objects have frequently been seen together," Smith writes, "the imagination acquires a habit of passing easily from the one to the other. If the first appear, we lay our account that the second is to follow. Of their own accord they put us in mind of one another, and the attention glides easily along them" (194). In the first account, imagination supplements the associative logic by providing tokens of things one never has experienced. In the second, it implements that logic by providing the glue between one idea and the next. We expect one idea to produce another because we remember that it has done so in the past. On the computational side, traces of ideas lodged in memory are reworked into the language of thought every time one has a new idea. As Fodor puts it, "you don't need an independent faculty of the imagination to implement inductive principles" because you can always feed traces into new thoughts: "records of X/Y coincidences are written in whatever language the mind computes in (Mentalese, say) and are stored in locations in the memory (for example, on the tape, assuming that the mind has the sort of architecture that Turing machines do). These records are themselves mental representation tokens; they are semantically interpretable and causally active and can be moved and copied, ad lib."[62] And so the object-image breaks down to the memory trace, whose meaning only inheres in its sentential function. Put next to Fodor, then, we can see how Smith tasks the imagination with joining whole pictures to one another and as a consequence of this function with producing pictures of things one never sees, like another person's mental state.

The point of comparing sympathy to computation is thus to reveal rather than to dismiss Smith's empiricism. Smith understood ideas to have the shape of pictures and thinking to be an association among them. Computers may not work that way. But novels might, or they might have. Many early works of fiction are of course quite concerned with coming up with forms to depict

thought. The question remains to what degree the account I have provided is useful to explain these forms. I have tried to show how eighteenth-century fiction and philosophy understood intentional states as relations to object-images and the process of attributing such states to oneself or others as a nesting of images within images. To put things this way is not to say that what novels really do is slow down or make explicit what is at the cognitive level an extremely fast and unnoticed process. Rather it is to say that novels represented thinking in such a manner and that novelists understood reading to occur in such a fashion. It is to say that thought often occurs in eighteenth-century fiction as a process of reciprocal image association also imagined to obtain in the process of reading. Such a version of thinking is both like and unlike what we see in contemporary work on the mind. Neither one has to be explained by the other in order see how models taken from the cognitive disciplines might sit in an interesting tension with the theory of the period itself. Both are evidently concerned with describing the way in which the mind works. And in the comparison, much of what is particular about thinking in eighteenth-century literature and philosophy comes sharply into view.

NOTES

1. See for example Kenneth MacLean, *John Locke and English Literature of the Eighteenth Century* (New Haven, 1936). For empiricism and the novel see, inter alia, Ian Watt, *The Rise of the Novel* (Berkeley, 1957), 9–34, and Michael McKeon, *The Origins of the English Novel 1600–1740* (Baltimore, 1987), 65–89. The particular question of mental architecture I discuss below is covered albeit in a slightly different context, by John Bender, *Imagining the Penitentiary: Fiction and the Architecture of the Mind in Eighteenth-Century England* (Chicago, 1987), 11–42. I will discuss several examples of cognitivist readings of fiction below. For something of an overview, see Alan Palmer, *Fictional Minds* (Lincoln, 2005).

2. Needless to say, philosophy of mind and cognitive science are not the same thing, nor without tremendous variation and internal debate. I concentrate here on the computational model, according to which mental processes are operations on syntactically structured symbols. Computation is particularly interesting because it shares with empiricism a representational theory of mind while differing on both epistemological and semantic grounds about what representation entails. See, for example, Jerry Fodor, *Psychosemantics: The Problem of Meaning the Philosophy of Mind* (Cambridge, 1987), especially 16–21. One opposed position is that perception provides unmediated access to the external world, associated in the eighteenth century with Thomas Reid (*Inquiries into the Principles of Com - mon Sense* [1764]) and in the contemporary philosophical scene with such "direct realists" as John McDowell (see his *Mind and World* [Cambridge, 1996]) and the later work of Hilary Putnam, who has argued that "our difficulty in seeing how our minds can be in genuine contact with the 'external' world is, in large part, the product of a disastrous idea that has haunted Western philosophy since the seventeenth century" (*The Threefold Cord: Mind, Body, and World* [New York, 1999], 43). Another opposed position, perhaps more important for the current essay, would be the related schools of "connectionism," which is in many respects a lineal descendent of empiricist associationism and "eliminative materialism," which attempts to reduce mentality to brain activity. For the former, see Gary Marcus, *The Algebraic Mind: Integrating Connectionism and Cognitive Science* (Cambridge, 2001); for the latter, Paul Churchland, *The Engine of Reason, The Seat of the Soul: A Philosophical Journey into the Brain* (Cambridge, 1995).

3. The "cognitive revolution" of the 1950s and '60s was in large part a reintroduction of the representational theory of mind against the then-dominant schools of behaviorism. Whereas behaviorism (so the story goes) tended to reduce mental states to actions, and in fact to argue that mental states as such were inscrutable and irrelevant, the cognitive revolution attempted to reason past actions to their antecedent beginnings in the mind. While it is true that one can never see a mind at work, cognitivists argued, it is also true that actions cloak the mental states by which they are shaped, or as Ned Block put it, that "intelligent behavior depends on the internal information processing that produces it" ("Psychologism and Behaviorism," *Philosophical Review* 90 no. 1, [1981], 5–43). If the classical statement of logical behaviorism was Gilbert Ryle's *The Concept of Mind* (New York, 1949), the classic statement of the cognitive revolution was Noam Chomsky's 1959 review of B. F. Skinner's *Verbal Behavior*: see *Readings in the Psychology of Language,* ed. Leon A. Jakobovits and Murray S. Miron (Englewood Cliffs, N.J., 1967), 142–55.

4. Thomas Hobbes, *Leviathan, or the Matter, Forme, & Power of a Common-Wealth Ecclesi - asticall and Civill,* ed. Richard Tuck (Cambridge, 1996), 13. Further references are to this edition and noted in parentheses.

5. Jerry Fodor, *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology* (Cambridge, 2001), 19. Fodor is particularly interesting for my current purposes because he is so involved with distinguishing computation from association and, as we will see, with establishing the relation between cognitive science and eighteenth-century philosophy.

6. Franz Brentano, *Psychology from an Empirical Standpoint* [1874] (London, 1995). It is a source of unfortunate confusion that "intentionality" means one thing in philosophy of mind and another in literary studies. Needless to say, literary studies has defined intentionality, since Wimsatt and Beardsley, largely in terms of motive, whereas for philosophy "intend" describes the mind's direction to an object and therefore marks the difference between mental and nonmental states. See Tim Crane, "Intentionality as the Mark of the Mental," *Contemporary Issues in the Philosophy of Mind,* ed. Anthony O'Hear (Oxford, 1994), 1–17.

7. In the past decade, Fodor has emphasized how little we know about domain-general or global computational properties of the mind in comparison with the progress made at the more local or modular level. A great deal of debate within cognitive science has turned on this question with some—like Steven Pinker, Leda Cosmides, and John Tooby—arguing against Fodor that the mind is "massively" modular (that there are, in effect, no domain-general capacities). This is especially relevant for the evolutionary question addressed in the second half of this essay. See Fodor, *The Mind Doesn't Work That Way,* 55–78.

8. All the writers I discuss in this essay are in one way or another materialists (Hobbes and Hume as well as Fodor and Goldman), in contrast both to the substance dualism of Descartes and the double-aspect monism of Spinoza (one substance, with physical and mental attributes that do not collide). For contemporary philosophy and cognitive science, the question is often posed in terms of supervenience: the mind supervenes on the brain—mental states are ultimately physical states—with varying degrees of abstraction between the one and the other. For a recent discussion see Jaegwon Kim, *Physicalism, or Something Near Enough* (Princeton, 2005), esp. 1–31.

9. One interesting feature of the computational theory of mind has been in fact the analytic separation of mental systems from the objects and entities in which they are implemented. According to the "multiple realizability" hypothesis elaborated by Hilary Putnam in the 1960s and '70s, for example, mental states could ostensibly be put into operation by machines as well as brains, aliens as well as humans. Following on this argument, one important dimension of the cognitive turn was to combine a functional account of mentality with a computational theory of mind, according to which thinking is performed in formal syntax by the kind of machines described by Alan Turing. In this respect, differences between empiricism and cognitive science amount to something like varia-

tions in operating systems. On multiple realizability, see "The Nature of Mental States" (408–28) and "The Mental Life of Some Machines" (429–40), both collected in Hilary Putnam, *Mind, Language, and Reality* (Cambridge, 1983). For the classical Turing model, see Alan Turing, "Intelligent Machinery," *The Essential Turing: Seminal Works in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life,* ed. Jack Copeland (Oxford, 2004), 395–432.

10. Locke's *Essay Concerning Human Understanding,* for example, frequently reminds us that we have little access to the inner structure of things whose "nominal essence" we construct by reflecting on experience. "I shall not," he says, "set my self to enquire philosophically into the peculiar constitution of bodies, and the configuration of parts, whereby they have the power to produce in us ideas of their sensible qualities." Ideas represent objects. They are also "objectively in the mind." But they are not, Locke insists, identical with the "things the mind contemplates" (*An Essay Concerning Human Understanding* [Oxford, 1975], book 2, chapter 21, section 75; dedication; 4.21.4.)

11. Locke, *Essay,* 2.12.2. Further references are to this edition and noted in parentheses.

12. I refer here to Fodor's "language of thought" hypothesis—that thinking happens in a non-natural language he dubbed "mentalese." See, for example, *The Language of Thought* (Cambridge, 1975) and the very useful appendix to *Psychosemantics,* 135–54.

13. See Fodor, "How the Mind Works: What We Still Don't Know," *Daedalus* (Summer 2006): 88.

14. See Zoltán Szabó, "Compositionality as Supervenience," *Linguistics and Philosophy* 23 (2000): 475–505.

15. Fodor, "How the Mind Works," 88.

16. Fodor, "How the Mind Works," 87.

17. Thus Fodor writes, "To a first approximation, then, the idea that there are mental representations is the idea that there are Ideas minus the idea that Ideas are images" (*Concepts: Where Cognitive Science Went Wrong* [Oxford, 1999], 8).

18. David Hume, *Treatise of Human Nature* (Oxford, 1978), 10.

19. In addition to the chapter of the *Treatise* noted above, see Hume, "On the Association of Ideas," *An Enquiry Concerning Human Understanding* [1748] sect. 3, ed. Tom L. Beauchamp (Oxford, 2000), 17–23.

20. Zenon Pylyshyn, *Seeing and Visualizing: It's Not What You Think* (Cambridge, 2003), 330.

21. Pylyshyn, 330.

22. Fodor, *Hume Variations* (Oxford, 2003), 134.

23. Fodor, *Hume Variations,* 134.

24. "Hume was right about his most fundamental architectural claim: there must be simple concepts and there must be mechanisms . . . that are able to construct complex concepts from them" (Fodor, *Hume Variations,* 83).

25. Fodor, *Hume Variations,* 95. The criticism of Hume's associationism recalls Kant, who argued that Hume "could not explain how it can be possible that the understanding must think concepts, which are not in themselves connected in the understanding, as being necessarily connected in the object, and since it never occurred to him that the understanding might itself, perhaps, through these concepts, be the author of the experience in which its objects are found, he was constrained to derive them from experience, namely, from a subjective necessity (that is from *custom*), which arises from repeated association in experience, and which comes mistakenly to be regarded as objective" (*Critique of Pure Reason,* trans. Norman Kemp Smith [New York, 1965], 127).

26. Fodor, *Hume Variations,* 97.

27. Fodor, *Hume Variations,* 133.

28. Fodor, *Hume Variations,* 144.

29. See Locke, *Essay,* 2.21. I discuss this chapter in "Locke's Desire," *The Yale Journal of Criticism* (Fall 1999): 189–208.

30. Fodor's position on consciousness is short and tart: "Nobody has the slightest idea

what consciousness is, or what it's for, or how it does what it's for (to say nothing of what it's made of)" ("You Can't Argue With a Novel," *London Review of Books* [March 4, 2004]: 31). His writing is dotted with such dismissals of the problem of consciousness, e.g., "I try never to think about consciousness. Or even to write about it" (*In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind* [Cambridge, 1999], 73). The point seems to be that computation is and ought to be silent on what features of mental process require an awareness of their occurrence.

31. The contemporary philosophical literature on consciousness and qualia is considerable, to say the least. For a basic discussion of phenomenal consciousness, see Block, "Some Concepts of Consciousness," *Philosophy of Mind: Classical and Contemporary Readings,* ed. David Chalmers (Oxford, 2002).

32. See Locke, *Essay,* 2.27.

33. For the debates over Locke's chapter on personal identity in early eighteenth-century literary intellectual culture, see Christopher Fox, *Locke Among the Scriblerians: Identity and Consciousness in Early Eighteenth Century Britain* (Berkeley, 1989).

34. Catherine Trotter, *A Defence of Essay of Human Understanding, written by Mr. Locke* (London, 1702), 29. Citations will appear hereafter in parentheses.

35. This is the argument made most famously by Hume's "bundle" theory of identity in the *Treatise,* a distant version of which can be found in Derek Parfit's *Reasons and Persons* (Oxford, 1984).

36. Trotter's novel was first published in Samuel Briscoe ed., *Letters of Love and Gallantry, and several other Subjects, All Written by Ladies,* vol. 1 (London, 1693). The passage here is from page 63, hereafter in parentheses.

37. The term "theory of mind" was coined by the primatologists David Premack and Guy Woodruff in their essay, "Does a Chimpanzee Have a Theory of Mind," *Behavioral and Brain Sciences* 1 (1978): 515–27, which discussed whether primates could attribute mental states to others (either primates or humans). The term then caught on among developmental cognitive psychologists and philosophers of mind, especially after Simon Baron-Cohen's popular crossover book, *Mindblindness: An Essay on Autism and Theory of Mind* (Cambridge, 1995), which presented in well-sculpted prose a decade's work of research connecting defective "mind reading" capacity to autism. "Theory of mind" now covers a lively debate about the nature of mind reading, especially whether the capacity works by theoretical inference (what is called somewhat inelegantly theory-theory) or by simulational enactment ("simulation theory"). Along with the theory-theory and simulation theory debate follows another about whether the capacity is domain-general (in the whole mind) or domain-specific (in a particular module). Baron-Cohen, for example, is a domain-specific theory-theorist because he claims that mind reading capacity works by a modular Theory of Mind Mechanism (ToMM), one that is informationally encapsulated (it receives no input from other parts of the mind, only outputs into the general processor), innate, and part of our evolutionary inheritance. ToMM comes "online" at roughly three or four years of age, at which point children begin to attribute mental states to others different from their own. For a recent discussion of the neuroscience of theory of mind, see Rebecca Saxe, "Why and How to Study Theory of Mind with fMRI," *Brain Research* 20 (2006): 57–65.

38. Alvin Goldman, *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading* (Oxford, 2006), 3.

39. Fodor, *Psychosemantics,* 3. Even if Berontus were to say to her "I am angry," Orinda would still be attributing a mental state of anger to him. She would be inferring that his statement that he is angry is in fact an indication of his having that feeling, and not, for example, a lie or an accident. As Fodor goes on to say, "inferring people's intentions from the sounds they make" includes a number of *ceteris paribus* clauses, from trustworthiness to not "being monolingual speakers of Urdu who happen to utter the words by accident" (4).

40. Blakey Vermeule, "Machiavellian Intelligence and Theory of Mind," introduction

to book in process, 28. I want to thank Vermeule for allowing me to quote from the manuscript. Lisa Zunshine, *Why We Read Fiction: Theory of Mind and the Novel* (Columbus, 2006).

41. Daniel Defoe, *Roxana: The Fortunate Mistress* [1724], ed. John Mullan (Oxford, 1996), 82.

42. Zunshine, 16, 22.

43. Zunshine, 82.

44. Zunshine, 100.

45. Thus the implicitly teleological form of the literary history: epistolarity, free-indirect discourse, stream of consciousness.

46. For the larger enterprise of which Baron-Cohen's *Mindblindness* is an important contribution, see note 37, above. It should be remarked that Baron-Cohen's book is a popularization of an academic debate within which he represents one (modular, evolutionary, and theoretical/inferential) but not the only position. For example, not all cognitive science or all nativism (e.g., Chomsky and Fodor) is evolutionary. For Fodor's critique of evolutionary psychology, see *The Mind Doesn't Work that Way,* especially 79–100—a response to Steven Pinker's *How the Mind Works* (New York, 1999)—along with his review of Pinker and Plotkin originally published in the *London Review of Books* and later collected in *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind* (Cambridge, 2002), 203–14.

47. Zunshine, 7

48. Baron-Cohen, 12.

49. I want to emphasize that I take no position (how could I?) on the evolutionary nature of mind reading, and that I want none of the unfortunate hostility literary study typically has to science in general and nativism in particular. Nevertheless, reverse engineering is one thing with respect to a known quantity, another with respect to highly debated terrain. We know enough about human anatomy, for example, to reason backward from the structure of the hand to the evolution of the prehensile thumb. Our knowledge of the mind, it seems from an even cursory look at the literature, is nowhere near as agreed-upon. Hence the reliance of evolutionary accounts of theory of mind on "just so" stories. What an advantage it must have been to be able to interpret behavior in terms of mental states! These are, of course, hardly questions for literary criticism to answer. My point is only that the evolutionary story leaves the structure of the mind an entirely settled phenomenon against which novels simply butt their heads.

50. The concept of modularity, central to many varieties of cognitive science, suggests that various parts of mind function are localized in particular domains or "modules." The important point is that modules are both domain specific (dedicated to one database, such as language function or mind reading), localized and hence subject to particular impairment, *and* informationally encapsulated, providing output (such as inferences about mental states, one's own and others') without receiving input from other areas of the mind. Hence, for example, the Müller Lyon illusion: we know the lines are of equal length yet that matters not at all to what we see. The classic work on modularity is, again, by Fodor, *The Modularity of Mind* (Cambridge, 1981). Since the publication of that book, Fodor has consistently chided his followers (like Pinker) that modularity only covers low-level aspects of cognition and that the mind must have some, still relatively unknown, domain-general capacity to compute over its various inputs. See, for example, *The Mind Doesn't Work That Way,* especially the discussion of abduction as a distinctly nonmodular process of thought, 41–54. ToMM was coined by Alan Leslie; see his "Pretense and Representation: The Origin of 'Theory of Mind,'" *Psychology Review* 94 (1987): 412–26. Baron-Cohen adds an Intentionality Detector (ID), an Eye Detection Detector (EDD), and Shared Attention Mechanism (SAM) to the list of modules. For a helpful overview of the debate about levels of modularity, see H. Clark Barrett and Robert Kurzban, "Modularity in Cognition: Framing the Debate," *Psychological Review* 3 (2006): 628–47. For a looser version of the modular account of mind reading, see Shaun Nichols and Steven Stich, *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding other Minds* (Oxford, 2003). Nichols

and Stich prefer the language of boxes and mental workspaces in part because it provides a way of understanding how agents create imaginary and possible scenarios about the mental life of others.

51. Not all theory-theory is modular, however. For the "child scientist" perspective favored by Alice Gopnik, theory of mind is domain-general and develops through a kind of naïve scientific method of experiment and failure in early psychogenesis. See e.g., Alice Gopnik, *Words, Things, and Theories* (New York, 1997).

52. Zunshine, 10.

53. Zunshine, 153–55.

54. Zunshine, 8.

55. Goldman, 4. See also Robert Gordon, "Folk Psychology as Simulation," *Mind and Language* 1 (1986): 158–71.

56. Goldman, 20.

57. Goldman, 151. For the discussion of modularity, see 95–112. Goldman suggests there may be modules for low-level mind reading of emotions but suggests that modularity has difficulty explaining the reading of propositional attitudes. Simulation theory has found support in the recent discovery of mirror neurons, so named because they reflect the activity of the neurons of targets. Agent P watches target S perform task T (typically a motion). The mirror neurons in P flash at the same brain location when S performs T even when P remains inactive. See Goldman, 132–36 and 202–20. Antonio and Hanna Demasio draw similar conclusions: "Explanations of the existence of mirror neurons have emphasized, quite appropriately, the role that mirror neurons can play in allowing us to understand the actions of others by placing us in a comparable body state. As we witness an action in another, our body-sensing brain adopts the body state we would have were we ourselves moving" ("Minding the Body," *Daedalus* [Summer 2006]: 18–25).

58. The connection between the subpersonal—whether neuro or cognitive—dimension and the necessarily more explicit and slow temporality of narration is something that is as yet underdeveloped in the cognitive approaches to literary study.

59. Adam Smith, *Theory of Moral Sentiments* (Bloomington, 1976), 9. Citations will appear hereafter in parentheses.

60. Goldman would call this low-level mind reading—the attribution of emotional states as opposed to propositional attitudes—in order to distinguish it from the high-level attribution of propositional attitudes. Simulation theory, on his account, holds a distinctive advantage to theory-theory in its conceptualization of what occurs at this lower level; see 113–46.

61. Fodor discusses Hume's concept of the imagination along these lines in *Hume Variations,* 114–33, but not in terms of the latter's concept of sympathy.

62. Fodor, *Hume Variations,* 130, 131.