Rationality, Intelligibility, and Interpretation

Peter Stone

1. INTRODUCTION

IN AN IMPORTANT SERIES OF BOOKS AND ARTICLES, Jon Elster has studied human rationality, both its scope and limitations and the central role it plays in social theory (see, e.g., Elster 1983b, 1984, 1989b). Elster firmly believes that the interpretation of human behavior could not be carried out without a theory of rationality. At the same time, however, he also believes that actual human beings are frequently irrational; indeed, he has amassed an extensive and growing catalogue of the irrationalities people routinely display. These two beliefs might seem to be in some sort of tension. If people are not rational at all, why worry about rationality when trying to make sense of their behavior? Why not treat rationality as one among many possible models of how people might behave—and a model with loads of disconfirming instances to boot (cf. Green and Shapiro 1994)? While Elster has on occasion attempted to explain why rationality is so indispensable to the interpretive project (e.g. Elster 1985), his efforts in this direction have been small as compared with his tireless devotion to identifying new types of irrationality and chronicling their occurrence.

In his latest paper, "Interpretation and Rational Choice," Elster distinguishes between *rationality* and *intelligibility* (Elster 2009, 7–9). This distinction is new to his work, and it is, I believe, a very useful one. Properly used, it allows one to explain the role rationality plays in the project of interpretation. It does so in a way that recognizes all of the limitations of rationality people display in the real world, and yet still places rationality at center stage. In this paper, I

Peter Stone is an Assistant Professor of Political Science at Stanford University. His areas of interest include theories of justice, democratic theory, rational choice theory, and the philosophy of the social sciences. He is currently conducting research on the role of lotteries in decision-making.

shall offer an account of rationality and intelligibility, and the roles they play in the interpretation of human behavior. This account draws extensively upon the work of Elster, both here and elsewhere. While Elster's own account of rationality and intelligibility differs somewhat from my own (in ways that I shall indicate where appropriate), I believe that my account is by and large compatible with Elster's previous work on rationality. It also clarifies how the process of interpretation works, whether employed by readers of fiction or by students of the social world. This clarification, in turn, has implications regarding what social scientists can and should contribute to the task of interpretation. Finally, it highlights an important similarity between the methodologies employed in the humanities and in the social sciences.

2. RATIONAL CHOICE

"We want," Elster declares, "to be rational" (Elster 1989b, 28). By this he means that people want their behavior to be governed by three optimizing operations. To satisfy these operations, they must act so as best to realize their preferences, given their beliefs. They must also adopt the best beliefs, given the information they have. And they must gather the best amount of information, given their beliefs and preferences (i.e. no large expenditures of energy in search of information regarding trivial decisions). An agent who performs these three operations successfully is to that extent a *rational* agent (Elster 2009, 3-4; see also Elster 1989b, 4).[1]

These operations impose various constraints on how an agent acts. It requires, for example, that one's preferences and beliefs satisfy various *consistency* conditions. An agent who is inconsistent—who believes claims that are logically contradictory, or whose preferences generate cycles (such that option x is preferred to option y, which is preferred to option z, which is preferred to option x)—cannot optimize successfully. I shall assume that these constraints are all met by a rational agent.[2]

This picture of rational behavior is, of course, greatly oversimplified. It leaves out elements that must surely be part of any fully developed theory of rational behavior. Rational action, for example, must surely be compatible with at least some forms of conformity to social norms, although how this might work has long been a subject of fierce contention (Elster 1989a). While there is much to say that could complicate the picture, the theory of rationality sketched thus far will be adequate to illustrate the relationship between rationality and intelligibility, and the relationship of both to the task of interpretation. I shall therefore not attempt to complicate the story here.

To say that people want to be rational is to state that people will, as far as they are able, endeavor to act in accordance with rationality's dictates. "We take little pride," Elster points out, "in our occasional or frequent irrationality" (Elster 1989b, 28). When a person comes to understand her behavior to be irrational, she usually tries to correct matters as best she can. (Indeed, one of the crucial roles that consciousness plays for human beings is to provide a means by

---

[1] Elster admits that there is more to rationality than these three operations. In particular, there are conditions of *autonomy* that one might wish to impose on preferences (Elster 1983b, chap. 1). Elster's embrace of autonomy, however, sits uneasily with his claim that "desires are the unmoved movers, reflecting Hume's dictum that 'reason is, and ought only to be the slave of the passions'" (Elster 1989b, 4). For my own argument that Elster's conception of rationality would benefit from a stronger connection with conditions upon preferences, see Stone (2003). For purposes of this paper, however, I shall treat Elster's trio of optimizing operations as the core of rationality.

[2] Elster describes the complete set of consistency conditions as dictating a *thin* theory of rationality (Elster 1983b, chap. 1). This theory places no constraints on the content of beliefs and preferences, only their form. The theory of rationality here adds further constraints to the content of beliefs

which they can survey what they are doing and make adjustments when necessary.)[3] Rationality is thus inherently a normative concept—indeed, one of the most essential such concepts, insofar as people constantly try to behave in accordance with its dictates.

To be committed to rationality, however, is not the same as being rational at all times. Human beings can fall short of rationality in their behavior in two distinct ways. First, they may lack the facilities necessary to behave in perfect accordance with what might in the abstract sound like the right way to behave. Real people, for example, lack *infinite sensitivity*; they cannot distinguish between outcomes that are sufficiently similar, even if real differences exist in principle. This can lead to violations of rationality. Most people would claim indifference between having no sugar in their coffee and having one grain of sugar, between having 1 grain and 2 grains, between having 2 grains and 3 grains, and so on. But they would not profess indifference between having no sugar in their coffee and having one million grains. The resulting lack of transitivity in judgments violates common understandings of how a perfectly rational actor would behave (Edwards 1954, 381).[4]

When it comes to rational behavior, *ought* most definitely implies *can*. Thus, one cannot fault oneself for failure to conform perfectly to an idealized standard of rationality, any more than a poor person should be faulted for buying the best food she can afford, even if that food is not very nutritious. An ideal picture of the perfectly rational actor, unhampered by any of the cognitive and other limitations real people possess, may have its uses, but for purposes of regulating their own behavior, people generally have a sense of the abilities and limitations they possess, and try to perform the three optimizing operations of rationality as best they can given those abilities and limitations. They try, in short, to *be as rational as they can*.[5]

They try, but they don't always succeed. This is the second way in which human beings can fall short of rationality in their behavior. Even when people possess the capacity to perform at a certain level of rationality, they do not always do so. People make mistakes, get careless, can be distracted, etc. And so they flub even activities well within their basic capacities, such as elementary mathematical operations. This fact, like the fact that human beings are not perfectly rational, should only be surprising to someone who believes that there are areas of human endeavor in which people have attained infallibility.

Thus, human beings by and large try to be rational. They know (or should know) that they will fall short in some respects due to the cognitive and other limitations they face, and that they must do the best they can given those limitations. They also know (or should know) that they will occasionally fall short even taking into account their limitations.[6] But by and large they will

---

[3] This is not to say that beings that lack a fully developed sense of consciousness (slugs, snails, etc.) are not rational in any way. It is, however, to say that it is impossible to imagine a conscious being that placed no value on the rationality of her behavior.

[4] Judgments of this kind can of course be rendered more sensitive through experience. A person might, through sufficient consumption of coffee with varying amounts of sugar, develop very fine-grained judgments as to precisely how much sugar she enjoys. But the investment of resources into learning experiences of this kind imposes opportunity costs on the actor, and past a point any such investment becomes a mistake. It is irrational, as Elster points out, to aspire to more rationality than one is capable of attaining (Elster 1989b, 107).

[5] The most obvious use for an ideal picture of rationality is as a yardstick, against which deviations from the ideal can be measured and compared.

[6] They may, however, have difficulty recognizing exactly when they have fallen short, or made a mistake. Indeed, this must be so, given that one of the mistakes to which people are prone is self-deception. Thus, people who are intellectually honest must admit that within a large class of their actions some of them must be irrational, while simultaneously believing each of these actions to be rational. This is similar to the problem of the author who believes

behave as rationally as they can. They will apply rationality as a regulative ideal. This claim is almost tautological; if a person consistently failed in a particular way, there would be every reason for her to call that failure a limitation, and to acknowledge herself as succeeding as best she can given that limitation. This does not mean that it is easy for a person to know her own limitations;[7] a person can fall into error both by aiming to be more rational than she can be and by allowing herself to be less rational than she can be. But if this were not the case, Reinhold Niebuhr's famous plea—"God, grant me the serenity to accept the things I cannot change, the courage to change the things I can, and the wisdom to know the difference"—would be trivially satisfied.[8]

This last point implies one further condition that an individual might impose on herself as a demand of rationality. The condition, simply put, is *know thyself*. Be aware, as far as possible, of one's own abilities and limitations. "As in Kant's first critique, the first task of reason is to recognize its own limitations and draw the boundaries within which it can operate" (Elster 1989b, 17). This condition can only be imposed on a rational actor capable of consciousness. An actor lacking consciousness cannot be held to this standard, but then again such an actor would be incapable of holding herself to any standard. As with every other rationality condition, a human being can only aspire to meet the demands of self-knowledge; self-deception is a common failing. But without this condition—a higher-order regulative condition, in effect—the effort to behave in accordance with the other conditions would come to naught.

Before moving on, it is important to note a fundamental ambiguity in the term "rationality" as used thus far. One can use the term to refer to how an actor with perfect cognitive powers would make decisions. Or one can use it to describe how an actor optimizes while making no mistakes subject to the real constraints she must observe. The boundary line between the two can be somewhat hard to draw; is an actor who lacks certain information about the world, but can draw perfect inferences from the information she has (in accordance with Bayes' rule) subject to constraints or not? But regardless of how the line is drawn, only the latter, but not the former, can serve as a regulative ideal for human beings who want to be rational. For this reason, I shall use the term "rationality" to refer to behavior that performs the three optimizing operations, without mistakes, subject to all existing constraints. But it should be understood as shorthand for "as rational as possible given the existing constraints." Thus, two individuals might both optimize according to their own constraints, but the first might face less serious constraints than the second. It makes sense, on my account, to say that both are behaving rationally (so long as they are making no mistakes), even though in a sense the first is more rational than the second.

This account of the normative role played by rationality is important, as it critically informs our efforts to interpret the behavior of others. It is to these efforts to which I now turn.

---

both that each individual statement in his book is correct, and that at least some of the statements in the book are incorrect. On the nature of this paradox, see Makinson (1965) and the discussion in Elster (1978, 88).

[7] Indeed, there may be an essential element of indeterminacy in this process. One might never be able to tell exactly where a borderline is without crossing over it. Cf. Elster (1984, 160). Similarly, one might never know whether a given level of rationality is within the scope of one's abilities without trying it out.

[8] The ideally rational actor thus strikes a balance between doggedly pursuing perfection and rationalizing away every failure. An actor who recognizes no constraints will waste time pursuing the unattainable; the actor who sees every failure as embodying a constraint will never accomplish anything. On this point see Statman (1987, 743).

3. INTERPRETATION, FICTION, AND NONFICTION

Equipped with an understanding of themselves as more-or-less rational, human beings must interact with their environment in order to survive, prosper, and achieve their goals. Within that environment, people encounter three types of entities, distinguishable by the different ways in which people interact with them. There are *intentional* beings, which pursue goals of their own. There are *designed* artifacts, which are constructed by other beings so as to realize certain goals. And there are *physical* objects that are neither intentional nor designed. Toward each of these classes of entities, people adopt different methods of interaction. To use the terminology of Daniel Dennett, from whom this three-part distinction is drawn, people adopt the *intentional stance* toward intentional beings, the *design stance* toward designed artifacts, and the *physical stance* toward physical objects (Dennett 1987).[9]

Human beings adopt one or more of these stances in interacting with the world more or less self-consciously. They also approach these interactions with more or less professionally developed intellectual tools. Thus, ordinary laypeople employ what Dennett has called "folk physics" to interact with physical objects, and "folk psychology" to deal with intentional beings. Scientists perform much the same tasks but, due to the advantages offered by the division of labor, are able to devote more resources to the intellectual tools within their area of specialization. They develop their physics or psychology or the like more systematically and self-consciously than laypeople, tightly constrained by the demands the world places upon them, can do. But the difference is of degree, not of kind. Scientific theories, on this view, are simply more carefully and critically developed versions of the intellectual instruments people employ in ordinary thinking.[10] There is therefore no principled gap between what the physicist and the layperson is doing with regard to physical objects, or between what the social scientist and the layperson is doing with regard to intentional beings. And so when I speak of someone adopting the intentional stance in order to cope with interactions with intentional beings, I shall have both professional and amateur efforts in mind.

With respect to each type of entity, the goal is the same—an explanation of why the entity does what it does. In the case of intentional beings, such an explanation is described as an *interpretation*, or an assignment of *meaning*. When such an explanation exists, the *explanandum* is thereby said to be rendered *intelligible*. Or, in Elster's words, "I shall understand interpretation as proposing and verifying hypotheses about *meaning*. Interpretation thus understood is a form of explanation rather than, as is often assumed, a mental operation that is contrasted with explanation" (Elster 2009, 5-6; Elster's emphasis).

In addition, an important part of the environment within which human beings interact involves fictional intentional agents, characters that exist only within the *stories* that other intentional agents have created. Human beings wish to explain the behavior of these agents as well.

[9] These categories, of course, are not mutually exclusive. An intentional being is also a physical object (barring the existence of any ghosts that may be out there). And one can try to interact with an intentional being purely *qua* physical object. This makes sense at times; if a person is thrown out of a window at some height directly above you, the smart thing to do is to react in the same manner as if that person were any other object of comparable weight. But the intentional stance is generally the most effective stance to take with respect to interacting with intentional beings. Indeed, the class of intentional beings (designed artifacts) is *defined* by reference to the class of entities with respect to which the intentional stance (design stance) is most effective (Dennett 1987, chap. 23). The physical stance thus emerges as the default category for when the other two prove unavailing.

[10] This is essentially a pragmatist view of human knowledge of the sort endorsed by Dewey (1910). For a more careful articulation and defense of this view, see Stone (2002).

They want to know why these agents do what they do, and for this purpose the intentional stance is also effective. This is true even though the goal of interpretation varies depending upon whether real or fictional agents are involved. In the case of the former, the goal is typically to make predictions as to how those same agents will behave in the future, and thus how one might best interact with them. In the case of the latter, the goal is typically what Elster calls "aesthetic" or "emotional" value (Elster 2009, 11); there is something compelling about watching the logic of a "good story" unfold. This is not all that is required for entertainment, but it is more or less a necessary condition; if one cannot make sense of what the characters in a story are doing, and why, then the story *qua* story is a failure. (Depending on the nature of the story, it might still provide entertainment due to gratuitous sex, gunplay, etc.)

The line between the purposes of interpreting real and fictional behavior is not as stark as I may have made it sound. To be sure, one can read fictional stories in hopes of drawing lessons as to how people really behave, so as to interact with them more effectively later. And one can read history in the manner of fiction as well. In that case, the desire for historical accuracy takes a back seat to the enjoyment of "a good yarn." In neither case is there a problem, so long as one realizes that (1) history written for enjoyment might not be history that provides an accurate assessment of what the characters it depicts were doing; and (2) compelling and enjoyable fiction need not depict how real human beings behave. But regardless of the purpose to which interpretations are to be used, the method for rendering the behavior intelligible remains the same.[11]

Thus, the methodology employed in the interpretation of behavior in fictional stories is basically the same as that employed in the interpretation of behavior in the real world. There is, however, one important difference between the two cases. In reading a fictional story, one can and must interpret both the behavior of the characters in the story *and* the behavior of the author who created those characters. This means that the audience simultaneously adopts the intentional stance towards these characters and also the design stance as well; the characters must behave intelligibly, and they must behave in such a way as to entertain (or shock, amuse, etc.) the audience. "Authors," writes Elster, "are under a double pressure: they need to make the plot move on, and to do so through intelligible actions and statements by the characters" (Elster 2009, 18). The work must meet other constraints to be a success—it must tell its story parsimoniously, the action "has to flow *downhill*, in the sense of minimizing the appeal to accidents and coincidences" (Elster 2009, 12; Elster's emphasis), etc.—but the requirement that both the author's behavior and the behavior of his creations be intelligible is nonnegotiable for any story *qua* story.[12]

---

[11] Richard Rorty claims at one point that we should endeavor to "see the social sciences as continuous with literature—as interpreting other people to us, and thus enlarging and deepening our sense of community" (Rorty 1982, 203). But whereas Rorty wishes to assimilate the effects social scientific studies can have on our behavior to the effects that novels can have, I am trying to assimilate how we read novels to how we construct social scientific studies, at least to the greatest extent that obvious differences between the two genres allow.

[12] Clearly, there is more to art than just telling stories. Many works of art do not endeavor to tell stories at all, but to impart artistic meaning or aesthetic enjoyment through other means. And in some forms of art, the storytelling is perfunctory, a frame upon which other forms of entertainment can be offered to an audience. (Ballet, opera, pornography, and professional wrestling all seem to work on this principle.) Nevertheless, to the extent that story matters, it must offer characters who behave intelligibly, where the demands of intelligibility are the same as in "real life." And even if storytelling is minimal or nonexistent, the demand that the *author's* behavior prove intelligible—that the work of art can be understood via the design stance—remains. Elster declares his focus to be on "pre-modern novels and plays guided by the constraint that events are presented *as if they could have been real*" (Elster's emphasis). But the method of interpretation appropriate to these stories applies equally to more fantastical stories, as in science fiction. And indeed Elster does contend that this "view of interpretation as explanation could in principle be applied to all art forms" (6).

"Like God," Elster writes, "the author is setting in motion a process in which each event can be explained twice over, first causally and then teleologically" (Elster 2009, 11). This comparison, of course, raises the possibility that real human behavior might be understood not only as the behavior of intentional agents, but also as the product of some higher-order design. If this were the case, one might (from some suitable vantage point) view the activities of the human race as a story playing itself out,[13] and evaluate and critique both how the characters in the story behave as well as how well the author makes use of them. This argument will appeal to creationists, but without convincing evidence of such an author, or what precisely she or he was endeavoring to do, the appeal to the design stance to evaluate human conduct becomes a form of functionalism, which has been ably critiqued by Elster elsewhere (Elster 1983a).[14] The distinction between fictional and nonfictional stories thus remains firm.

4. INTERPRETATION AND RATIONALITY

Thus far, I have emphasized the similarity between interpreting behavior in fiction and doing so in the real world (including both formal social science and informal daily activity). I have not, however, described how precisely this task is performed. What does one *do* when one takes up the intentional stance in trying to explain what an entity does? This section addresses itself to this question.

The quest to render the behavior of others intelligible begins with a simple question. What would *I* do were I in the same situation? Can I see myself performing the action that that person performed, under a similar set of circumstances and constraints? If I can, then the action is intelligible to me.[15]

The task of rendering a person's behavior intelligible begins with the fact that when one is the interpreter, one perceives oneself to be a more-or-less rational actor acting under a set of constraints, both external (e.g. physical environment, resources) and internal (e.g. cognitive limitations). As interpreter, one projects oneself into the circumstances of the subject whose behavior is to be interpreted. This means that the interpreter expects the subject to behave more-or-less rationally under the set of constraints he faces. If the set of constraints for both interpreter and subject is identical, then the interpreter expects the behavior of both to be the same. Where this is not the case, however, the interpreter must decide how the set of constraints faced by the subject differs from her own set of constraints, and base her interpretations upon this decision. The move, however, is away from the set of constraints of the interpreter, which is the

---

[13] A particularly poetic example of this story is told by Mephistopheles to Faust. See Russell (1957).

[14] Daniel Dennett insists that it is intelligible to speak of intentions without an intentional actor, and thus to examine human behavior from the design stance (with "mother nature" as the designer) as well as the intentional stance. His argument, however, implies that all designed artifacts may be viewed from the intentional stance; a Coke machine, for example, could be described as "wanting to give you a Coke in exchange for several quarters," where the origins of this desire lie in the design humans used in creating the machine. Dennett embraces this implication, and uses it to criticize those who would distinguish entities with "real" intentions (human beings) from those with "derived" intentions (Coke machines). This move, however, would appear to collapse the design and the intentional stances into each other, a conclusion incompatible with Dennett's own typology of explanatory approaches. See Dennett (1987, chap. 8).

[15] Dennett distinguishes between a *Normative Principle* of interpretation, which assumes that another actor will do what it ought to do under the circumstances; and a *Projective Principle*, which assumes that another actor will do what the interpreter would have done under those circumstances (Dennett 1987, 342-43). In practice, the two principles are hard to distinguish (*ibid.* chap. 10; see also chap. 8). To the extent that the two principles do diverge, the argument offered here aligns itself with the Projective Principle.

natural starting point; where there is no reason for presupposing a difference, the interpreter assumes the set of constraints to be the same for both. (The assumption that the subject will be as rational as possible given those constraints, however, is not negotiable. The interpreter tries to be rational, and therefore attributes the same efforts to the subject.[16])

In practice, an agent engaged in interpretation employs a rich set of rules to decide what capacities and constraints agents different from herself possess.[17] A woman might assume, for example, that if she receives a confidential note from her boss—a boss not prone to loose lips—she can proceed as if other agents do *not* know the contents of the note.[18] She might also assume that a professional in a highly technical field, like physics or engineering, will be proficient in mathematics, even if the interpreter herself is bad at math. In the former case, the interpreter assumes the interpreted faces a tighter set of constraints than she does; in the latter case, the set of constraints is assumed to be looser, so that a "higher" degree of rationality is possible.[19] It may be hard for a social scientist (or literary critic) interested in interpretation to identify how an interpreter goes about assigning constraints that vary from her own.[20] But this merely suggests that the exploration of how agents do this could yield large dividends in terms of understanding the interpretive process.[21]

In practice, of course, people are only more-or-less rational; they do make mistakes. And an interpreter who puts herself in the place of a subject must recognize that this subject, even taking into account the constraints under which he must work, may fail to be rational. It is not possible *ex ante* to predict that a mistake will be made; if it was, then the resulting failure of action will

---

[16] Note that the interpreter must attribute both preferences and beliefs to a subject in order to make sense of that subject's behavior. A given set of beliefs by itself might be compatible with an infinite variety of behavior patterns, given the right preferences. Thus, I believe that the interpreter must attribute to the subject preferences, as well as beliefs, that are similar to her own as a baseline for interpretation. This is a controversial claim, one that I can only note in passing here. Elster himself denies it, but does not offer an alternative account as to how interpreters are to identify preferences.

[17] Interpretation does not require an explanation for the capacities and constraints. A person might be terrible at making mathematical calculations due to an identifiable brain impairment. Once the impairment is identified, that person's behavior is perfectly intelligible. How the agent got the impairment might be an interesting question, but it does not fall into the realm of interpretation proper. And when it comes to explaining capacities and constraints, one is usually moving away from the intentional stance and towards the design or physical stance (unless of course an intentional agent influenced those capacities and constraints).

[18] It takes time for human beings to develop the ability to make assumptions like this. A child under the age of five, for example, has great difficulty with the idea that other people may have false beliefs, that other people may not know what the child knows (Saxe 2004). The fact that children across cultures acquire this ability more or less at the same time suggests that many of the rules people employ in constructing interpretations are determined by the cognitive structure of the human mind, much as with language. I would like to thank Joshua Cohen for drawing my attention to this area of cognitive research.

[19] Once an interpreter attributes a tighter or looser set of constraints to a subject than that she herself faces, behavior that might have been intelligible *ex ante* might be harder to understand *ex post*. If the interpreter is bad at math, she will find another person's mathematical failings perfectly intelligible. But if she learns that the other person is a mathematician, those same failings may become unintelligible, at least until some other constraint is identified. (Perhaps the mathematician was drunk at the time.)

[20] The average person employs some such set of rules with little effort, a set that generally works very well. (Interpretive failure is the exception and not the rule.) This person, however, would have great difficulty articulating the set of rules so employed. Again, the comparison with the rules of language people employ is instructive here.

[21] Cf. Daniel Dennett on the process of identifying limiting conditions governing the behavior of other agents: "When do we—or must we—stop adding conditions? There is no principled limit that I can see, but I do not think this is a *vicious* regress, because it typically stabilizes and stops after a few moves, and for however long it continues, the discoveries it provokes are potentially illuminating" (Dennett 1987, 264; Dennett's emphasis).

look less like a mistake and more like a constraint (albeit a constraint that the agent failed to observe).[22] However, a mistake may still be intelligible *ex post* if the interpreter can imagine making the same mistake under those conditions. This is as good as the intentional stance can do when it comes to handling mistakes. (The mistake itself might still be explicable, but only in terms of the design or physical stance.)

Not all mistakes satisfy this condition, of course. A man does not mistake his wife for a hat, barring truly extraordinary impairments of some kind (Sacks 1998). Mistakes that fall outside the range of the interpreter's experience are difficult to grasp, and when they can be grasped they usually require recognizing that the subject faces constraints that the interpreter had hitherto thought impossible. Still, learning that certain types of mistakes are possible for others is a distinct part of the interpretive process, just as learning that certain types of constraints might be faced by others is.

5. RATIONALITY AND INTELLIGIBILITY

We are now in a position to talk about rationality, intelligibility, and their relationship in the practice of interpretation. A subject's behavior, on my account, is *intelligible* to an interpreter when that interpreter can see herself behaving the same way, given the constraints under which the subject was functioning. It is *rational* to the extent that the subject performs the three basic optimizing operations with respect to actions, beliefs, and information. Given that interpreters ordinarily aspire to be rational, and perceive themselves to be rational most of the time (except for occasional mistakes), this means that interpreters will find behavior intelligible to the extent that they can articulate either a set of constraints under which the behavior becomes rational, or else identify an understandable mistake the subject might be making.[23]

Interpretation, on this argument, requires the assumption of rationality—not the assumption that people will reason flawlessly all the time, but that people will typically reason as best they can given their constraints, while making an occasional mistake along the way. Interpretation requires this of us because it is what *we* try to do in our own behavior, and because interpretation is all about putting ourselves in the shoes of others. We have no choice when interpreting but to assume that others are doing what we would do, *mutatis mutandis*, in similar situations.

This conclusion is largely compatible with that drawn by Elster. The latter is, however, subject to possible misinterpretation because of Elster's choice of emphasis. In his paper, Elster stresses the difficulties of employing the rationality assumption, rather than the indispensability of this assumption. This is evident when he points out that "if we assume rationality" while interpreting the behavior of real human beings, "the behavior may appear more opaque rather than less" (Elster 2009, 5). He asserts this because people sometimes lie, misrepresent their in-

---

[22] Shaquille O'Neal may not be a very good free throw shooter, but one must assume that when O'Neal steps up to shoot, he is endeavoring to put the ball in the basket, subject to the constraints that his skill level imposes. If O'Neal— or an interpreter placing himself in O'Neal's very big shoes—could predict his failure in advance, then his behavior would be unintelligible in those terms (although it might be intelligible in some other terms, like *pretending* to make a free throw). Nobody performs an action intending to make any mistake, much less a particular type of mistake, at least without further qualification.

[23] People aspire to be rational, and view themselves as being so by and large. Almost trivially, this means that their behavior is normally intelligible to themselves. But not always. A person may fail to grasp or misunderstand her own cognitive limitations, or make mistakes she does not recognize. The first case would include behavior dictated by unconscious "drives," of the sort that psychoanalysis is supposed to reveal. The second case would include a patient with Alzheimer's disease, whose deteriorating condition leads to mistakes (e.g. leaving the alarm clock in the freezer) that the patient "just doesn't make."

tentions, etc.[24] Elster is correct to point out the difficulties that deception can cause for the interpretative project. But he cannot, consistently with the remainder of his argument, claim that the assumption of rationality itself can impede the interpretation of behavior. If he believed this, he would have to specify an alternative basis upon which interpretation could proceed that would somehow circumvent the problem of deception, and this he never does. And in the end, Elster takes for granted the possibility that an interpreter might see through misrepresentations so long as there really is some fact of the matter lying behind the deceptions (Elster 2009, 24). But in figuring out what this fact of the matter might be, interpreters must assume that their subjects are behaving more or less rationally. This implies that rational behavior might involve deception, but this poses problems only at the level of application, not of theory. Every agent has lied at least once, and had good reasons for doing so at least once. When an agent puts herself into the shoes of a subject whose behavior she is trying to interpret, then, she should have no problem recognizing deceptive behavior *per se* as intelligible. (Determining *when* deception is taking place, of course, is another matter.)

Rationality and intelligibility are thus closely related in the practice of interpretation, although they do not always travel together. Before concluding, it therefore makes sense to consider the four possible ways that these two attributes may combine or fail to combine together.

Behavior can be *rational* and *intelligible*. Indeed, when the project of interpretation is going well, this will be the normal state of affairs, for the reasons already given. And behavior can be *irrational* and *unintelligible*. To the extent that an interpreter cannot make sense of the subject's behavior as either the product of the three optimizing operations—operations she herself would strive to perform if she was in the subject's shoes—or as a mistaken departure from these operations, the subject's behavior doesn't really look like behavior at all. The intentional stance is not working in such a case—although this does not prove that it cannot work—and so perhaps the design or physical stance would be more appropriate. But these are the easy cases, the cases where rationality and intelligibility coincide. More interesting are the other two possibilities, the possibilities in which they diverge.

Can behavior be *irrational* but *intelligible*? Clearly. People make mistakes; as long as the mistakes made by a subject are mistakes the interpreter might recognize herself as making, the subject's behavior is intelligible (but irrational) to the interpreter. Weakness of the will and self-deception both lead to irrational behavior, as Elster notes, but such behavior is intelligible to anyone who has ever succumbed to temptation or temporarily deceived herself about her own motivations (Elster 2009, 8-9). Note that an interpreter might have a hard time explaining *why* she makes a particular mistake, but this need not affect the intelligibility of the mistake, whether it be committed by herself or by others. Elster makes this point using the example of Othello and Desdemona. Othello's worst fear is that Desdemona is unfaithful, and it is this fear that leads him to believe that she is unfaithful. Such irrational belief formation is the opposite of wishful thinking, and it seems hard to understand why anyone would be prone to such a mistake. According to Elster, such a mistake must involve the physiological equivalent of wires getting crossed somewhere (Elster 2009, 13; see also 8-9). But Othello's mistake is still a recognizable one for

---

[24] Elster implies that this is only a problem when it comes to interpreting the behavior of real human beings, not fictional ones. It is unclear, however, why this restriction should hold. Surely, fictional characters are capable of lying, misrepresentation, etc. And authors (who are, after all, real human beings) may sometimes craft their stories with an eye to deceiving the reader, at least temporarily. Thus, if the behavior of fictional characters can be explained twice over, deception can figure into both explanations.

most people; his behavior is therefore irrational but intelligible. Further understanding of why he is prone to such a mistake—why and how the wires got crossed—requires reference to either the design or the physical stance.

As the case of Othello and Desdemona makes plain, irrational but intelligible behavior can also be seen in fiction, and it influences the two perspectives from which we read stories. As Elster puts it, "*we expect the author to be rational and the characters to be intelligible*" (Elster 2009, 10; Elster's emphasis). We want the author to tell us the best story he can, subject to the constraints he faces. And part of that means that he gives us intelligible characters. And so "we may blame him if his characters fail to be intelligible. But we do not blame him simply because the characters fail to be rational, except if their irrationality is 'out of character'" (Elster 2009, 11).

As for the author, his work might fall short in several different ways, all of which should be intelligible to most people. Moreover, each of these shortfalls will likely generate a different audience reaction when recognized. As Elster points out, "Authors…are under a double pressure: they need to make the plot move on, and to do so through intelligible actions and statements by the characters" (Elster 2009, 18). The author may not perform both tasks at once. In that case, either "causality" will be sacrificed to "teleology" (i.e. the action in the story is unintelligible, but the plot moves to a climax and resolution), or vice versa (i.e. the characters' behavior makes sense, but the story goes nowhere). Either way, it could be the case that the author had the talent to produce a good story, but made several mistakes along the way. A novelist who regularly produces excellent novels will likely be forgiven a dud or two along the way. Such duds are the result of intelligible but irrational behavior. Less forgivable are authors for whom failings like this stem not from mistakes, but from constraints. These authors simply lack the talent to create intelligible characters and tell a good story with them. They may be doing their best, but their best isn't very good. Such authors may be rational in a sense—they produces the best stories they can—and yet still blameworthy. We don't just want the best story an author can produce; we want a good story, and if the best story an author can deliver isn't very good, then he should probably consider another line of work. (To the extent that the author fails to recognize his own lack of talent, he is guilty of a failure of rationality, albeit an intelligible one.) Worst of all is when an author is capable of writing a good story, but doesn't even try. Such behavior is intelligible and even rational if the author is simply maximizing according to another set of goals—for example, putting as little effort into writing as possible while still getting paid (Elster 2009, 18). This is perhaps the worst sin an author can commit against his audience. Moviemakers who make the best films they can despite a total lack of talent—Ed Wood is the classic example here—may attract and keep a fan following for years, while films made "for a paycheck" by people with far more talent are usually quickly forgotten.[25]

By far the most interesting scenario involves behavior that is *unintelligible* but *rational*. This could only happen in the case of a subject who optimizes within constraints, aside from the occasional mistake, but where either the constraints or the mistakes are opaque to the interpreter. In other words, there is a perspective from which the subject's behavior is rational, but the interpreter does not see it.

The most likely way in which this could happen is if the subject of interpretation is, to put it crassly, smarter than the interpreter. That is, if the subject faces fewer constraints on rationality

---

[25] This lesson applies in real life as well. We are willing to forgive people their occasional missteps, but are more prone to hold people accountable when they bite off more than they can chew, and even more prone to do this in the case of outright negligence.

than the interpreter, or makes fewer and less serious mistakes, it might be very hard for the interpreter to understand the situation. Granted, any non-megalomaniacal interpreter knows that there are people who can do things that she cannot do. When she has good reason to believe she has encountered one of them, she will attribute to that subject abilities that she does not have. But it might be hard for her to recognize exactly what kinds of problems the subject can solve, or even when the subject has in fact solved one of them.[26] If I ask a friend to handle my taxes for me because I believe he understands tax law better than I do, I can only hope that he can recognize a problem in my taxes that he cannot solve (and is nice enough to tell me about it). Otherwise, I may not discover the limits of my friend's abilities—abilities that are clearly less restricted than my own in this area—until I get audited.[27]

In fiction, this problem arises whenever an author depicts a character who is supposed to have abilities and skills that his audience will by and large lack. How does one convey credibly that the character has such abilities? The weakest way is simply to *tell* the audience about the fact, without worrying about whether the character actually displays those abilities at all. To do this is to take full advantage of the inability of most audience members to distinguish an authentic character with those abilities from an inauthentic one. Unfortunately, this approach can make the story laughable for an audience member who *has* the requisite skills, and recognizes that the character is not at all behaving realistically. Any game theorist watching the film *A Beautiful Mind* (2001), for example, will most likely cringe at how little one sees of the fantastic mathematical skills attributed by everyone to the character John Nash (and possessed by the real John Nash), and how the little one does see is generally off the mark (Landsburg 2002).[28] An online movie reviewer invented the term "informed attribute" to refer to a desirable ability a filmmaker attributes to a character solely by verbal description ("Why, he's one of the most brilliant biologists in the world!");[29] needless to say, he did not intend the term to be complimentary. And the phenomenon he diagnosed with this term is common enough in cinema that many web denizens have made use of the term as well.

Some works, however, manage to convey exceptional abilities without recourse to the dreaded informed attribute. The mystery stories of Arthur Conan Doyle, for example, remain popular after a century because the character of Sherlock Holmes is both brilliant and intelligible. Holmes carefully explains how he reached his conclusions, to the point where even the skeptical Dr. Watson recognizes them as "elementary." But the obviousness is completely *ex post*;

---

[26] This problem arises in evolutionary theory, as natural selection may lead an organism to have abilities whose function biologists have a hard time recognizing. Hence what evolutionary theorists call Orgel's Second Rule: "Evolution is cleverer than you are" (Dennett 2005).

[27] This problem arises in some accounts of political representation. If I appoint an agent to act on my behalf, and I believe the agent has the same abilities that I do, then I shall evaluate the agent's behavior by asking if he did the same thing I would do on my own behalf in the same situation. (I set aside the usual questions of how principles can successfully monitor agents, etc.) But if I appoint an agent who has abilities I lack—an appointment I presumably make *because* I lack these abilities—then I certainly don't want the agent to do the same thing I would do on my own behalf. If I was bad at math, and as a result I hired a tax attorney, I certainly wouldn't want him to break down crying at the sight of all those numbers (Pitkin 1967, 145). Under these circumstances, however, achieving accountability, as any theory of representation would demand, might prove very challenging.

[28] Similar complaints can be made whenever the motion picture industry depicts a brilliant scientist or mathematician, as in *Good Will Hunting* (1997), *Proof* (2005), or *Pi* (1998). For an extended complaint on this convention, see Stone (2007).

[29] See Jabootu's Bad Movie Dimension, "Glossary of Specialized Terms for Jabootu's Dimension of Bad Movies," http://www.jabootu.com/glossary.htm.

while it takes no brilliance to see a mystery's solution after it has been explained, it took brilliance to recognize the solution in the first place. Holmes thus manages to be smarter than his average reader without thereby leaving any of them in the dark as to how his highly rational mind works. (Something similar happens in the play *Copenhagen* [Frayn 2000], which features Neils Bohr and Werner Heisenberg discussing quantum physics at a level that non-quantum physicists can understand.)

One specific failure of rationality an interpreter might demonstrate is a failure of self-awareness. As interpreter, one might fail to recognize oneself as acting under certain constraints, or making mistakes of a particular kind. When confronted with a subject (fictional or not) facing similar constraints and making similar mistakes, an interpreter might find that subject's behavior unintelligible, even though it is just as rational as her own. The subject *does* exactly what the interpreter would do under similar circumstances, but the interpreter does not *recognize* this to be the case. One example of this phenomenon is offered by Elster. Most people tend to attribute greater cross-situational consistency to their own behavior than is warranted. But if an author offers a character whose behavior at any given time is governed by the particular situation at hand, then that character may be *both* as rational as the typical audience member *and* yet unintelligible to that member (Elster 2009, 14-15). A more troubling example might be that of a bigot. Take as example a woman who abhors the ignorant stereotypes employed by a fictional bigot from another racial or ethnic group, oblivious to her own use of essentially identical stereotypes. The case would be different if the real bigot agreed with the reasoning of the fictional bigot. In this case, the interpreter would find the subject both intelligible and rational, completely missing the fact that both interpreter and subject are making serious mistakes of belief formation.[30]

In the example just given, if the real bigot abhors the fictional bigot's behavior, she would presumably alter her own behavior if she became aware of the way art imitates life. This generates the possibility that an author can reveal to an audience member her own mistakes, of which she might not be aware, and thereby help her to become more rational. Mark Twain's *The Adventures of Huckleberry Finn*, for example, offers a teenage boy who shares all the racial prejudices of the antebellum South in which Twain was raised. Through contact with a runaway slave named Jim, with whom he forms an intense bond, Finn comes to recognize that his beliefs regarding Jim are unjustified. This process takes place in scenes like the following:[31]

---

[30] Both of these cases are different from the examples of unintelligible-but-rational behavior offered by Elster—Euripides' Medea and Racine's Phèdre. Elster believes that each character is "lucid about her self-destructive passions" to an implausible degree (12). In each of these cases, the subject has more self-knowledge than the interpreter. But the self-knowledge seems to be of the sort that the interpreter does not simply lack, but which the interpreter does not think is possible. An agent in the grip of such self-destructive passions should not be able to recognize these passions after calm reflection; one or the other has to give, either through the cooling of the passions or the creation of self-deceptive beliefs. For this reason, this is not properly a case of behavior that is unintelligible but rational, but of behavior that is rational only on the surface—of what possible use could calm-but-helpless reflection on self-destructive passions be?—and in a way that seems unintelligible. People exert at most the same degree of rationality in their examinations of their behavior as they do in the behavior itself; the former are never more rational than the latter.

[31] This is not to claim, of course, that Finn's beliefs on the subject of race become fully rational by the end of the novel. Finn does not generalize those beliefs beyond Jim to all slaves (Sidnell 1967) and has difficulty understanding those beliefs as anything other than a rejection of all morality (Bennett 1974). Nevertheless, Finn's new beliefs could hardly represent anything other than an improvement—not simply with respect to morality, but with respect to rationality as well—over the old.

> When I waked up, just at day-break, he [Jim] was setting there with his head down betwixt his knees, moaning and mourning to himself. I didn't take notice, nor let on. I knowed what it was about. He was thinking about his wife and his children, away up yonder, and he was low and homesick; because he hasn't ever been away from home before in his life; and I do believe he cared just as much for his people as white folks does for their'n. It don't seem natural, but I reckon it's so.
>
> <div align="right">(Twain 1996, 201)</div>

In Finn, Twain offers a character that a racially prejudiced audience would find intelligible. That character eventually comes to recognize his own prejudices (regarding Jim, at least) and overcome them, thereby becoming more rational in his beliefs. Twain might have hoped that his audience would come to a similar recognition through their encounter with Finn.

6. INTERPRETATION AND RELATIVISM

As these examples remind us, intelligibility is relative to the interpreter. What might be intelligible to an interpreter who recognizes one set of constraints and mistakes might be unintelligible to an interpreter recognizing different constraints and mistakes. Different interpretations will rationally arise from different starting points using the intentional stance. It is even plausible to suppose that people from a common culture may perceive in their shared experiences certain behavior as intelligible where those from a different culture might not. But one must not make too much of this seemingly relativistic point. For a given set of constraints and mistakes, the intelligibility of a certain set of behavior is an objective fact. Anyone who cannot understand the behavior in terms of that set is making a mistake. The existence of a set of constraints and mistakes facing a given actor is similarly objective, even though the actor and his interpreters might both fail to recognize them. (For either interpreter or actor, such a failure would count as a mistake, or possibly as a constraint if sufficient information about the set of constraints and mistakes is unavailable.) And while interpreters from different cultural backgrounds might see people as facing different constraints or making different mistakes, the demands that rationality imposes on interpretation are common to all. Any interpretation that does not amount to rational maximization under constraint—by attributing to people cyclic preferences or logically inconsistent beliefs, for example—must either count such failure as a mistake, or else be rejected as violating the need for intelligibility.

It follows that all starting points for interpretation are not created equal. If different interpreters perceive certain people as facing different sets of constraints, one may simply have identified the real set of constraints more accurately than the other. More commonly, two different interpreters may each have a piece of the truth, and the understanding of human abilities and limitations that results from synthesizing their perspectives is better than that of either original perspective.[32] And the adoption of this synthesized understanding by interpreters cannot help but impact their understanding of their own abilities and limitations. The successful interpreter, after all, views others as behaving by and large as she would behave, given the constraints they face. If others are perceived to face constraints that the interpreters did not recognize before, she

---

[32] Cf. Dennett (1987, 29): "one can even acknowledge the *interest relativity* of belief attributions and grant that given the different interests of different cultures, for instance, the beliefs and desires one culture would attribute to a member might be quite different from the beliefs and desires another culture would attribute to that very person. But supposing that were so in a particular case, there would be the further facts about *how well* each of the rival intentional strategies worked for predicting the behavior of that person" (Dennett's emphasis).

must consider the possibility that she faces similar constraints herself. This is especially so if the subjects of her interpretation are or were unaware of some constraint; if they missed it, is she doing the same?[33]

If interpretation depends on a working model of how a rational actor behaves under various alternative sets of constraints (and possible mistakes), and if the superiority of some such models over others is a matter of objective fact, then the development of superior models is of critical importance to the interpretive project. Social scientists can play a crucial role here. Just as physicists, thanks to the division of labor, are able to develop models of the physical world superior to the "folk physics" of everyday understanding, so can social scientists improve upon the models embedded in "folk psychology." Social scientists that recognize this fact will best be able to contribute to actual interpretive practice.

At its best, social science makes just this kind of contribution. Consider, for example, the question of revolutionary social change. For decades Marx and his followers faced the problem of explaining why workers failed to undertake revolutions that were (in Marxist eyes) clearly in the interests of the proletariat. This problem was frequently understood as one of explaining irrational behavior, of figuring out why the masses were continually doing the wrong thing. But the discovery of the collective action problem (and its two-person analogue, the Prisoner's Dilemma) by social scientists changed matters. It was now clear that, even if a course of action was in the collective interest of the proletariat, it need not be in the individual interest of any particular proletarian (Buchanan 1979). Workers, in other words, have *reasons* for not risking everything by hitting the barricades—reasons that can make an observer say, that's what I'd do if I were in their shoes. These reasons were not clearly recognized by political commentators (including Marx) before the collective action problem fleshed out the theory of rational action in new ways. Social science made the unintelligible intelligible in this case.

Ordinarily, there is a period of delay between the discovery of scientific results (be they in the natural or the social realm) and their dissemination throughout the general public. Nevertheless, the dissemination takes place, and when it does it changes the way people understand themselves and others. It thereby changes their behavior. And it changes the way in which people read and write. Interpreters who become more sophisticated in their interpretations, by adopting more sophisticated and rich models of rationality, will no longer find intelligible the same behavior that they once did. Fictional characters that at one time might seem plausible become at a later time easy to dismiss, because "people simply don't behave like that." Fiction, therefore, must keep pace with popular understandings of what counts as intelligible and rational behavior; and what counts as intelligible and rational behavior will change under the impact of social science.

7. CONCLUSION

The relationship between social science and the humanities has proven difficult in the modern era (Snow 1993). For some, the two fields are orthogonal to each other, with different topics and different methods, such that no collaboration between them would make sense. For others, the fields deal with similar topics (like understanding human nature) but with methodologies that have nothing in common and that are mutually exclusive. The reflections on rationality and intelligibil-

---

[33] This is, I believe, the correct way to understand Charles Taylor's famous advice to interpreters: "in the sciences of man in so far as they are hermeneutical there can be a valid response to 'I don't understand' which takes the form, not only 'develop your intuitions,' but more radically 'change yourself'" (Taylor 1985, 54.).

ity offered here by both Elster and myself hopefully suggest that there are real similarities of method between the humanities and social science. Both are centrally concerned with the interpretation of behavior, and both generate interpretations in much the same way. All of this suggests that further informed collaboration between the two fields would prove very fruitful. | A |

ACKNOWLEDGMENTS

REFERENCES

Landsburg, S. 2002. Mindless. *Wall Street Journal*, February 22.

Makinson, D.C. 1965. The paradox of the preface. *Analysis* 25:205-7.

Pitkin, H. 1967. *The concept of representation*. Berkeley: University of California Press.

Rorty, R. 1982. *The consequences of pragmatism*. Minneapolis: University of Minnesota Press.

Russell, B. 1957. A free man's worship. In *Why I am not a Christian*, ed. P. Edwards, 104-16. New York: Simon and Schuster.

Sacks, O. 1998. *The man who mistook his wife for a hat*. New York: Touchstone.

Saxe, R. 2004. Reading your mind: How our brains help us understand other people. *Boston Review*, February/March, 39-41.

Sidnell, M. J. 1967. Huck Finn and Jim: Their abortive freedom ride. *Cambridge Quarterly* 2:203-11.

Snow, C. P. 1993. *The two cultures*. Canto ed. New York: Cambridge University Press.

Statman, M. 1987. Review of *Sour Grapes*, by Jon Elster. *Journal of Economic Literature* 25:742-43.

Stone, P. 2002. Review of *Microfoundations, Method, and Causation: On the Philosophy of the Social Sciences*, by Daniel Little. *Philosophy of the Social Sciences* 32:120-26.

———. 2003. The impossibility of rational politics? *Politics, Philosophy and Economics* 2:239-63.

———. 2007. Pi and the movie mind. *Philosophy Now*, November/December, 44-46.

Taylor, C. 1985. Interpretation and the sciences of man. In *Philosophy and the Human Sciences: Philosophical Papers*, 2:15-57. New York: Cambridge University Press.

Twain, M. 1996. *Adventures of Huckleberry Finn*. New York: Oxford University Press.