Interpretation and Rational Choice[*]

Jon Elster

INTRODUCTION

THE INTERPRETATION OF TEXTS AND OF BEHAVIOR is closely related to the rational choice explanation of action. In this article I try to bring out two aspects of that connection. I first ask whether we can enhance our understanding of a literary text by assuming the rationality of its creator. In novels, for instance, things often happen "for a reason"—be it a reason for the author, for the characters, or for both. The assumption of rationality can remove opacity. I then ask whether, when we try to account for behavior, the rationality of the agents might not be an obstacle to explanation. When they claim to be motivated by this or that aim, they may do so "for a reason" that is self-serving rather than truth-seeking. If we assume rationality, the behavior may appear more opaque rather than less.

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.[1] Some readers might accept that claim when ap-

This article was originally published in *Rationality and Society*. DOI: 10.1177/1043463108099347 *Rationality and Society* 21 (2009): 5

Jon Elster holds the Chair of Rationality and Social Science at the Collège de France. His books include *Ulysses and the Sirens* (1979), *Sour Grapes* (1983), *Making Sense of Marx* (1985), *The Cement of Society* (1989), *Solomonic Judgments* (1989), *Nuts and Bolts for the Social Sciences* (1989), *Local Justice* (1992), *Political Psychology* (1993), *Alchemies of the Mind* (1999), *Ulysses Unbound* (2000), *Closing the Books: Transitional Justice in Historical Perspective* (2004), and *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences* (2007). His research interests include the theory of rational choice, the theory of distributive justice and the history of social thought. He is currently studying the history of constitution-making processes, as well as the microfoundations of civil war.

* Presented at the Conference on Rational Choice Theory and the Humanities, Stanford University, Stanford, CA, April 29–30, 2005 (http://www.stanford.edu/group/RCTandHumanities/).

[1] Max Weber (1969, 33) wrote, for instance, that natural science does not aim at "understanding" the behavior of cells. Føllesdal (1979) offers an account more in line with the one proposed here.

plied to behavior, but reject the application to texts. I do not see any alternative, however, if we want to avoid subjective and arbitrary readings. In fact, for that reason I believe that the view of interpretation as explanation applies to all art forms. Except for one example from the visual arts, however, I only discuss novels and plays. In fact, I limit myself further to pre-modern novels and plays guided by the constraint that events are presented as if they could have been real. As I shall argue, an analysis of classical literature allows us to differentiate the category of meaning more finely into authorial rationality and character intelligibility.

On the explanation-based view of interpretation, we assume that there is a fact of the matter and that the aim of the scholar is to get it right. It may be hard to decide which to prefer among several interpretations of a given work, but we know that at most one of them can be true. The same is true with regard to determining the motives of historical actors. However difficult, the task makes sense because there is, usually at least, a fact of the matter. In fact, for agents to try to hide their real motivations there has to be something for them to hide. Some mental states may, to be sure, be so fluid, unstable or uncrystallized that no verbal statement could adequately capture their content. (As the French moralist Alain remarked, "I can form a mental image of the Parthenon, but not count the columns.") Such cases provide no grounds, however, for wholesale skepticism.

If interpretation is explanation, it has to be causal explanation, since there is no other kind. It is subject to the general principles that guide and constrain causal explanation, such as the need for the *explanans* to be temporally prior to the *explanandum*. Even functional explanation, in which we explain events by their consequences, is subject to this rule. When we come across a valid functional explanation the *explanandum* is never a token, always a type. If the consequence of an event occurring at one point of time makes it more likely for similar events to occur at later times, we can say, in shorthand, that the event is explained by its consequences. A child may initially cry simply because he or she feels pain, but if the crying also gets attention from the parents the child may start crying for attention as well. As we shall see, however, the metaphysical absurdity of trying to explain token-events by their consequences has not prevented scholars from trying to do so.

RATIONALITY AND INTELLIGIBILITY

Figure 1 presents what I take to be the standard view of rationality. The aim of rational-choice theory is to prescribe what it is rational for an agent to do in a given situation and to explain the agent's behavior by the hypothesis that he or she follows the prescription. As indicated in the diagram, rational choice is defined in terms of the relations among action, desires, beliefs and

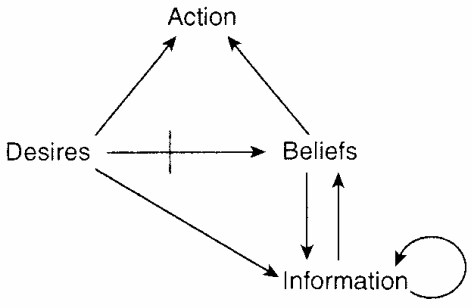

Figure 1.

information. Minimally, rational-choice theory must tell the agents how best to realize their desires, given their beliefs. Furthermore, the theory must prescribe which beliefs it is rational for the agents to hold, given their evidence or information. Needless to say, a belief may be rational and yet false. Finally, since rational agents cannot always simply limit themselves to the evidence already at hand, the theory must tell them how much new information to acquire before they form the beliefs on which they can base a decision.

We see that belief rationality can be shaped by the agent's desires, not directly (as in wishful thinking), but indirectly through the process of information acquisition. For a person who does not care much about the future there is no point in reading consumer reports about the durability of various brands of dishwashers. The optimal investment in information acquisition also depends, as shown by the loop, on the information acquired in the early stages of the search. Suppose I am out in the woods plucking berries. I know that berries tend to come in clusters, so I am prepared to spend some time looking before I start plucking. If I strike lucky and find an abundant patch right at the beginning, I would be foolish to keep on looking.

If an action is rational, it is usually also intelligible, except that we might sometimes find it hard to understand how the agent could remain rational in the given circumstances. Irrational behavior can also, however, be intelligible. I shall distinguish among three varieties of intelligible but irrational behavior that are particularly relevant for my purposes here, and contrast them with some cases of unintelligible behavior.

The first arises when the machinery of decision described in Fig. 1 is truncated in one way or another. By virtue of its peculiar urgency, a strong emotion may prevent the agent from "looking around" (i.e., gathering information) before acting. Rather than taking his time, the agent rushes into action without considering the consequences ("marry in haste, repent at leisure"). Another form of truncation arises in weakness of will, understood as acting against one's own better judgment. (I shall question this understanding later.) The person who has decided to quit smoking yet accepts the offer of a cigarette acts on a reason, namely a desire to smoke. For an action to be rational, however, it has to be optimal in light of the totality of reasons, not just one of them.

A second variety arises in the short-circuiting of the machinery of decision. As I indicate in Figure 1, there need not be anything objectionable in desires having an indirect effect on belief formation, mediated by information-gathering. However, a direct causal impact (represented by the blocked arrow in the diagram) is unacceptable. A subtler form of motivated belief formation arises when the agent stops gathering information when the evidence gathered so far supports the belief he or she would like to be true. In either case, the outcome is a biased belief. (By contrast, the truncated form only generates low-quality beliefs.) These forms of motivated belief formation are, in their way, optimizing processes: they maximize the pleasure the agent derives from his or her beliefs about the world (rather than the pleasure he or she can expect from his or her encounters with the world). Sometimes such beliefs are held merely for their consumption value, but they can also serve as premises for action. The process is constrained, in the sense that the agent cannot completely discard long-standing and well-established beliefs just because holding them is aversive. Smokers may not be able to persuade themselves that smoking is harmless, but might revise the risks downwards in a biased way.

A third variety is what we might call a wire-crossing in the machinery of decision, to use a phrase coined by Amos Tversky (in conversation). We can easily understand why the mind might engage in cognitive dissonance reduction (wishful thinking being one variety), but why, Tversky asked, should it also pursue dissonance production? A French proverb says, "We believe easily

what we hope and what we fear." The second half is the puzzle: why would fear of a bad outcome make us see it as more likely than is warranted by our evidence? If the belief is supported neither by the evidence nor by our desires, why adopt it? Clearly, nothing is being optimized.

Nevertheless the behavior is intelligible because it arises from the belief–desire system of the agent. It is slightly less intelligible, perhaps, than actions arising from truncation and short-circuiting, but more so than behavior that seems completely isolated from that system.

Actions that elude interpretation include those caused by compulsions and obsessions, phobic behaviors, self-mutilations, anorexia and the like. To be sure, these behaviors have the effect, which explains why they are performed, of relieving the anxiety the agent feels if he does not perform them. Yet washing one's hand fifty times a day or walking up fifty flights of stairs to avoid the elevator is not like taking a tranquillizer. Taking Valium may be as rational and intelligible as taking aspirin,

but compulsive and phobic behaviors are unintelligible because they are not part of an interconnected system of beliefs and desires. Or to take an example from John Rawls, we would find it hard to understand the behavior of someone who devoted his time to counting blades of grass unless it was linked to some other goal, such as winning a bet.

The belief of a disturbed man that the dentist in the building next door is directing X-rays at him to destroy his mind is unintelligible. By contrast, in politics paranoid beliefs are intelligible because they are rooted in the desires of the agent. A strongly anti-Semitic person is motivated to entertain absurd beliefs about the omnipotent and evil nature of the Jews. It's not that he wants Jews to have these features, but he is motivated to believe they do because the belief can rationalize his urge to destroy them. Even contradictory beliefs may be intelligible. An anti-

Semite may on different occasions characterize the Jews as "vermin" and assert their omnipotence. The very same people who say, "Jews are always trying to push in where they are not wanted" also believe that "Jews are clannish, always sticking together." One and the same Muslim may assert that the Israeli intelligence service Mossad was behind the attacks on the World Trade Center on September 11, 2001, and take pride in the event.

INTERPRETING WORKS OF LITERATURE

We may illustrate some of these ideas by a classical, almost trite problem in literary criticism: why does Hamlet delay taking revenge for his father's death? Many explanations have been offered. Some of them appeal to irrationality, in terms of weakness of will or clinical depression. There is, however, also a simple rational-choice account. Although Hamlet initially believes what his father's ghost told him about Claudius, he later decides to gather more information by staging a play to "catch the conscience of the king." Once the reactions of the king have confirmed his belief, however, he lacks an opportunity to realize his desire, which is to make Claudius burn in hell forever. Although he has the occasion to kill Claudius while he is praying, doing so would according to contemporary theology bring him salvation rather than damnation. Later, he kills Polonius behind a curtain, falsely but not irrationally believing him to be the king. As Edward Wagenknecht says, "As it turns out, it was an unfortunate move . . . but it was not a foolish move. . . . We must judge Hamlet on the basis of the information he possessed at the time his act was committed."[2] We may add that on this occasion, he had no reason to gather more information,

[2] Wagenknecht (1949, 193).

since he could reasonably assume that someone hiding behind the curtains in the queen's presence would be the king.

I do not claim that this is the right interpretation. Indeed, up to this point I have not said anything about what it would mean for an interpretation to be "right." My point is simply that the three episodes I have mentioned are prima facie consistent with the idea that Hamlet is rationally pursuing the goal of avenging his father's murder. Another question is whether it is consistent with Hamlet's repeated self-accusations of lacking the resolve to take revenge. Many commentators interpret these as a sign of weakness of will, and the two first episodes as based on self-deceptive excuses for inaction. (The third episode is harder to square with this view.) Now, although weakness of will and self-deception violate the canons of rationality, they are, I have argued, perfectly intelligible.

Before I proceed, let me quote another insightful observation by Wagenknecht:

> The primary, the most important reason why Hamlet does not kill the king sooner is that the death of the king must involve the end of the play. Fundamentally this consideration has nothing to do with psychology. It is simply a matter of dramatic necessity. The nature of the revenge play was such that no other plan of action was possible. The hero received his commission at the beginning: He did not act upon it until the end.[3]

Clearly, Shakespeare could not have had Hamlet say "I cannot kill the King until Act V." By contrast, towards the end of Ibsen's *Peer Gynt*, when Peer is afraid of drowning, the strange passenger tells him "one does not die in the middle of the fifth act." One of the conventions of the classical stance, respected by Shakespeare[4] but not (in *Peer Gynt*) by Ibsen, is indeed that the characters are portrayed as if they could have been real. Given this convention, we expect the author to be rational and the characters to be intelligible. If the author fails to be rational, in the sense I shall explain, we blame him or her. As a special case, we may blame the author if his or her characters fail to be intelligible. But we do not blame the author simply because the characters fail to be rational, except if their irrationality is "out of character."

Authorial rationality is like the rationality of God. Like God, the author is setting in motion a process in which each event can be explained twice over, first causally and then teleologically. I take this idea from Leibniz, who wrote that there

> are like two kingdoms, one of efficient causes, the other of final, each of which separately suffices in detail to give a reason for the whole, as if the other did not exist. But neither is adequate without the other when we consider their origin, for they emanate from one source in which the power that makes efficient causes, and the wisdom which rules final causes, are found united.[5]

---

[3] Ibid., 189.

[4] Empson (1982, 85) argues that Hamlet's reference to his father's ghost as "this fellow in the cellarage" (a technical term referring to the area under the stage) is "a recklessly comic throw-away of an illusion," similar in that respect to the remark of the strange passenger. I cannot claim that a non-Pirandello reading is superior, but assume it for the sake of argument.

[5] Leibniz (1969, 588). For an explicit comparison between the creation of a work of art and God's creation of the world, see also Cannon (1971, 220): "Man stands in the same double relationship to his world and to his Creator as a character in a play to the scene of action and to the playwright."

God's aim is to create the best of all possible worlds. Specified to include the temporal dimension, the idea can be understood as the best of all possible sequences.[6] Although the transition from one state of the universe to the next occurs by ordinary physical causality, the initial state and the laws of causality have been chosen so as to maximize the overall perfection of the sequence.

If we limit ourselves to the classical drama or the classical novel, the author's task is to develop the plot through what the characters say and do, often in response to one another. His or her aim is to do so in a way that maximizes the aesthetic and ultimately the emotional value an intended reader will derive from the work.[7] Thus each action or statement by a character can be explained twice over,[8] as a reaction to previous actions and statements (or external events) and as a generator of surprise, tension and ultimately tension resolution in the reader. The first explanation rests on intelligibility, the second on rationality, in a sense I shall now try to clarify.

The fact that authors often make many drafts before they are satisfied, or before they lay down their pen, is evidence that they are engaged in a process of choice and that they possess—but may not be able to state—criteria for betterness. Often these drafts involve small variations, suggesting that the authors are aiming at a local maximum of whatever form of betterness they are striving for.[9] However, the difference between an author and someone who is merely climbing along a gradient is that the former's creativity goes beyond mere choice. The reason why the creation of a work of literature cannot be reduced to rational choice is that the number of meaningful word sequences is too large for one person to scan them all and select "the best."[10] Although a "rational creator" may try to make the problem more tractable by deliberately excluding some sequences,[11] too many will usually remain for choice to be a feasible selection mechanism. Instead, the author will have to rely on his or her unconscious associative machinery.

---

[6] See Elster (1975, 216–17) for two occurrences of the phrase "la plus parfaite de toutes les suites possible" in Leibniz. In each, he notes that at any given moment, even in the best of all possible sequences, the world might be less perfect than it could have been.

[7] In Elster (2000, 105–9), I argue that in the final account the value of a work of art consists of its capacity to generate emotions, including weaker and purer forms of the emotions we experience in everyday life as well as the specifically aesthetic emotions of wonder, amazement, surprise, humor, relief, and release.

[8] "Suppose we want to know 'why' in the early part of Dickens's *Great Expectations* . . . the six- or seven-year old Pip aids the runaway convict. Two different kinds of answer are possible: (1) according to the logic of verisimilitude (made prominent, in fact, by the text): the child was frightened into submission; (2) according to the structural needs of the plot: this act is necessary for Magwitch to be grateful to Pip so as to wish to repay him; without it the plot would not be the kind of plot it is" (Rimmon-Kenan 1983, 17–18).

[9] This is not to deny that there can be large differences among successive drafts. Yet even when one of these is chosen as the final version, there is usually some fine tuning to be done.

[10] Incidentally, as Dagfinn Føllesdal (1982) points out, many applications of rational-choice theory to more ordinary choices also fail to predict behavior because they do not and cannot take into account the power of the imagination to come up with new and hitherto unthought-of alternatives. Rational-choice theories of technical change based on the idea of an "innovation possibility frontier" fail for the simple reason that the economic agents have no way of knowing where that frontier is located (Elster 1983, 105). And as Humphrey Lyttleton is reported to have said, "If I knew where jazz was going, I'd be there already."

[11] This is the topic of Elster (2000, chap. 3). Although some self-imposed constraints are intended to make the choice easier, others are meant to make it harder. The unities of space, time and action illustrate the former case. The latter is illustrated by a novel by Georges Perec, *La disparition*, in which the letter "e" is nowhere used. By eliminating the most frequently used letter in the language, Perec made it impossible for himself to resort to the lazy and self-indulgent language that can be the bane of even good novelists, thus implementing Mallarmé's program of "donner un sens plus pur aux mots de la tribu." The use of rhyme and meter in poetry has the effect (among others) of concentrating the mind in a similar way.

Rational creation is therefore largely about getting the second decimal right or, to shift the metaphor, about climbing to the top of the nearest hill. In yet a further metaphor, this is a left-hemisphere task.[12] The right-hemisphere task of getting the first decimal right, or finding a hill that towers over the others, is not within the scope of rationality. Yet even reduced to the task of fine-tuning, authorial rationality matters. It may be better to find the top of a small hill (a "minor masterpiece") than to remain on the slopes of a larger one (a "flawed masterpiece"). Let me enumerate and then discuss some demands that rationality imposes on the author. First, the acts and utterances of the characters have to be intelligible. Second, the author has to meet the twin requirements of fullness and parsimony. Third, the work has to flow downhill, in the sense of minimizing the appeal to accidents and coincidences. Fourth, it has to offer a psychologically gratifying pattern of the build-up and resolution of tension.

Intelligibility can be absolute or relative and, if relative, global or local. The question of absolute intelligibility is whether any human being could ever behave in this way. The question of relative local intelligibility is whether the behavior of a fictional person is consistent with his or her behavior in similar situations earlier in the work. Whereas the requirements of absolute and of relative local intelligibility are crucial constraints on authorial rationality, that of relative global intelligibility consistency across situations is not. If anything, the respect for the latter constraint may be seen as an aesthetic flaw.

In some cases, absolute intelligibility may be violated by excess of rationality. Consider for instance Euripides's *Medea*. When she is about to kill her children, she says, "I know indeed what evil I intend to do. But stronger than all my after thoughts is my fury." In Ovid's version, Medea says, "An unknown compulsion bears me, all reluctant down. Urged this way or that . . . I see the better and approve it, but I follow the worse" (*video meliora proboque, deteriora sequor*). Racine's *Phèdre*, too, is equally lucid about her self-destructive passions. *Medea* and *Phèdre* are portrayed as being subject to weakness of will in the strict sense, knowing that what they are doing is contrary to the all-things-considered judgment they hold at the very moment of acting. Although passion causes them to deviate from that judgment, it does not affect it. Racine's Hermione in *Andromaque* is a more credible character. Her judgment being clouded by her emotions, she is self-deceptive rather than weak-willed. My suggestion—it is nothing more than that—is that the simultaneous presence of extreme emotion inducing extreme behavior and full cognitive lucidity goes against what we know about human nature.

Whereas too much rationality can be unintelligible, irrationality can be perfectly intelligible. What can be more intelligible than the reaction of M. de Rênal in Stendhal's *Le Rouge et le Noir* when, in the face of strong signs that his wife is having an affair with Julien Sorel, he chooses to believe in her fidelity? The wish is the father of the thought. More paradoxical are cases in which the desire that one's wife be faithful causes the belief that she is not, against the evidence. In Othello, "Trifles light as air are to the jealous confirmation strong as proofs from holy writ." The first is a case of short-circuiting, the second one of wire-crossing.

The distinction between global and local relative intelligibility may be illustrated by a remark by Knut Hamsun. After 1945 Hamsun, who had collaborated with the Nazis during the war, underwent psychiatric observation to determine whether he was mentally capable of being tried (he was 86 years old at the time). When the psychiatric professor asked him to describe his "main character traits," he replied as follows:

[12] Although the left brain–right brain distinction has become a metaphor for the analytical–creative distinction, matters seem to be more complicated (Bekhtereva et al. 2000).

> The so-called naturalistic period—Zola and his time—wrote about persons with main character traits. They had no use for nuanced psychology. People had one dominant capacity that governed their actions. Dostoyevsky and others taught us all something different about people. From the very beginning I do not think there is a single person in any of my writings with this dominant and unitary capacity. They are all without so-called character—they are divided and fragmented, not good not bad, but both. Nuanced and changing in their mind and in their actions. This is no doubt how I am myself. It is very possible that I am aggressive, and that I have a little of the other traits the professor suggested—vulnerable, suspicious, selfish, generous, jealous, righteous, logical, sensitive, a cold nature. All these would be human traits, but I cannot give any of them the preponderance in myself.[13]

Hamsun's claim matches arguments by psychologists that character traits tend to be local rather than global.[14] Folk psychology, as summarized in many proverbs, tends to attribute greater cross-situational consistency to people than their behavior actually displays. Except for psychopaths, people rarely display lack of compassion in all situations. A person may, for instance, be consistently indifferent towards beggars and consistently compassionate towards sick co-workers. As Proust wrote, "one might have thought" that the young men in *Le temps retrouvé* paid for inflicting pain on the customers of Jupien's brothel were "fundamentally bad, but not only were they wonderful soldiers during the war, true 'heroes,' they had just as often been kind and generous in civil life."[15]

Like Zola, Balzac has also been criticized for portraying individuals as near-automata in the grip of a "dominant and unitary" passion.[16] In their defense one might say that such people exist, and perhaps they do. Yet if this becomes an author's standard mode of showing the characters, the adherence to global relative intelligibility may end up violating absolute intelligibility. Since many readers probably share the illusion that people are all of one piece, this flaw may easily go unnoticed. If authors transcends folk psychology and allow their characters to behave differently in different situations, they may disappoint their readers except "the happy few" for whom they are really writing. Hamsun cites Dostoyevsky, and one might also mention Stendhal.[17]

Either a Balzac or a Stendhal may be tempted to have a person act "out of character," according to their respectively global or local conceptions of what it means to act in character. An author may paint himself into a corner, so that the only way to develop the plot as planned is to allow for one character to act in an inconsistently impulsive or hesitant manner. Yet, as Wagenknecht says, "if Hamlet delays only to prolong the play, then the play is badly constructed." Shakespeare must "convinc[e] us that Hamlet is a man who, under the given set of cir-

---

[13] Langfeldt and Ødegård (1978, 82).

[14] Ross and Nisbett (1991); Doris (2002); Mischel (2004).

[15] Proust (2003, 145).

[16] Fernandez (1981, 62, 64).

[17] The much-discussed question whether it is plausible for Julien Sorel to shoot Mme de Rênal when she denounces him to his employer (and the father of his pregnant mistress) must thus be answered in the affirmative. Clearly there is nothing absolutely intelligible about his behavior. Relative (local) intelligibility is vacuously satisfied, since he had never been in a similar situation before.

cumstances, would behave quite as he does behave."[18] A plot should develop like water seeking its natural downhill course, not by the author forcing it to run uphill.

Let me illustrate this idea by some of Stendhal's marginal comments in the manuscript of his unfinished and posthumously published novel *Lucien Leuwen*. Stendhal has the eponymous hero fall in love with a young widow, Mme de Chasteller. His feelings are reciprocated, but he does not dare to reach out to her. The very delicacy of mind that makes him superior to "the most accomplished Don Juan" and hence capable of inspiring love also makes him inferior to any "less well-bred young Parisian" who would instantly know how to handle the situation. To move the plot forward, Stendhal needs to bring them together, but does not quite know how to do it. He writes in the margin: "Upon which the chronicler says: one cannot expect a virtuous woman to give herself absolutely; she has to be taken. The best hunting dog can do no more than bring the game within gunshot. If the hunter doesn't shoot, the dog is helpless. The novelist is like the dog of his hero."[19] The comment strikingly illustrates the need for the behavior of characters in a novel to be "in character."

Stendhal does eventually manage to engineer a situation in which the love of Lucien and Mme de Chasteller for each other can be shown and understood, and yet not be declared. But his difficulties do not end there. Stendhal's plan for the novel followed the dialectical Hollywood recipe: boy meets girl, boy and girl break up, boy and girl reunite. As we just saw, he had problems getting the thesis established. To produce the antithesis, Stendhal uses the ridiculous and manifestly teleological device of making Lucien believe that Mme de Chasteller, whom he has seen daily at close quarters, has suddenly given birth to a child. But what really stumped him was the synthesis. Although we do not know why he never got around to writing the third part in which the lovers would be reunited, one conjecture is that their union would not be plausible. In the second part of the novel, after the breakup, Lucien turns into a bit of a cynical rake, fundamentally honest by the lax standards of the July Monarchy but certainly very different from the awkwardly delicate person with whom Mme de Chasteller had fallen in love. Stendhal may have decided that if he had her love the transformed Lucien, it would violate relative intelligibility.[20]

As I understand the act of writing, the author is in a situation of Gricean communication with his readers or audience. "The role of the artist, properly understood, requires the artist, in the creation of his or her work, to adopt or bear in mind the role of the spectator."[21] More specifically, I have in mind Grice's two axioms of Quantity: "(1) Make your contribution as informative as is required for the current purposes of the exchange. (2) Do not make your contribution more informative than is required."[22] Aristotle comes very close to making the same requirement in the *Poetics*.[23] The two axioms correspond in fact directly to the aesthetic ideals of fullness and parsi-

[18] Wagenknecht (1949, 189). Referring to the conjectural Ur-Hamlet by Thomas Kyd, Empson (1982, 82, 89) writes that "it was thought absurdly theatrical … because it kept the audience waiting without obvious reason in the characters."

[19] Stendhal (1952, 1537).

[20] Alternatively, he could have let Mme de Chasteller undergo a similar Bildung, so that when she and Lucien meet again they can laugh affectionately at their earlier selves, as do Elizabeth Bennet and Darcy when the scales fall from their eyes. In his novels, however, Stendhal did not give his heroines that kind of development.

[21] Budd (1995, 11).

[22] Grice (1989, 26).

[23] "[T]he story . . . must represent one action, a complete whole, with its several incidents so closely connected that the transposition or withdrawal of any one of them will disjoint and dislocate the whole. For that which makes no perceptible difference by its presence or absence is no real part of the whole" (1351a, 31–35).

mony. The reader is entitled to expect that if the author says that it was raining when a character left a house, it is because the premise of rain will be needed later on, or to believe that a speech attributed to a character is intended to tell us something about the person.

In the aesthetic context, the axioms have to be broadly interpreted, since the aim of a work of art is not merely (or maybe not at all) to convey information. Rather, the aim that the reader is entitled to impute to the author is that of producing a local maximum of aesthetic satisfaction. Thus redundancy is not always to be eschewed, since it can serve an aesthetic function. To convey boredom, redundancy might be more effective than a mere authorial statement. Yet even then, there would come a point where the repetition would bore the reader rather than evoking the boredom of the character.

Conversely, potentially relevant details may be deliberately left out, to leave some room for the imagination of the reader. Rational creation is compatible with leaving blanks to be filled out by readers or viewers, but if artists overestimate the imagination of their audiences their efforts will be deemed a failure.[24] Suppose a novelist tries to suggest the temperamental incompatibility of the hero and heroine by having the street numbers of the houses in which they live being mutually prime, that is, with no common divisor. Barring special circumstances, he cannot count on readers being able to pick up that meaning. Moreover, any reader who does pick it up is likely to read non-intended meanings into all sort of other features of the work. The mere fact that a text is consistent with some numerological pattern does not entitle us to infer that the author was aware of it and intended his readers to perceive it, any more than data-mining and curve-fitting in the social sciences entitle us to believe that the patterns they uncover have causal significance.

To digress for a moment from fiction, we may take an example from the history of "unfinished prints," some of which may have been left deliberately unfinished and thus were, in a sense, finished. Commenting on a self-portrait by Van Dyck, showing only the head and the vaguest outline of the body, Peter Parshall writes that it was "the first instance we have of a consciously important print that was intentionally distributed by the artist in an unfinished state" and goes on to say that

> Although rare, proofs of the self-portrait survive in sufficient number to assume that Van Dyck himself must have deemed it ready for distribution, perhaps selectively at first. One of the earliest recipients appears to have been the humanist, connoisseur and art patron, Constantin Huygens, who composed an epigram, "On Van Dyck's Self-Portrait," that envisions the artist's intention: "Van Dyck, many depict a face, eyes/No one a matchless hand." These terse lines honor Van Dyck with a paradoxical conceit ornamented with a pun. Huygens interprets the absent portion of the figure as a declaration of Van Dyck's inimitable talent, a hand that can be represented only through its art. In Van Dyck's sparse self-image absence is invested with a very particular and present meaning. Being the purest record of Van Dyck's draftmanship, the unfinished proof gives the better account of the hand that created it.[25]

---

[24] Davidson (1993, 305–6) argues that "the intention by the originator that an utterance or writing be *interpreted* in a certain way is only a necessary condition for that being *the correct interpretation*; it is also necessary that the intention be reasonable" (my italics). By a reasonable intention he means that the author could reasonably assume that the intended reader would understand him in the intended way. If we see interpretation as a mode of explanation, however, the intention to be understood in a certain way is sufficient for the correctness. Davidson seems in this passage to conflate subjective intention and objective success.

[25] Parsall (2001, 19).

If the intention of the artist was to convey this *présence en creux* and if one highly qualified viewer understood that to have been his intention, it must be counted an artistic success. If he had no such intention, it would still be an artistic success because of the inimitable draftsmanship, but without that additional dimension. If he had the intention, and if no one then or later[26] had perceived it, it would to that extent count as a failure.

Earlier, I referred to the "downhill" character of a good plot, using acting "in character" as an example. More generally, good plots should not turn on unlikely events, accidents and coincidences. In *Middlemarch*, the encounter between Raffles and Mr. Bulstrode—a crucial element in the development of the story—is so contrived that it detracts from the otherwise seamless progression of the novel. Accidents can, to be sure, have their place in a novel. The accidental death of a parent may trigger or shape the unfolding of a plot, as may the death of both parents in the same accident. But if the plot requires their deaths in two separate accidents, credulity is strained. The convenient death of a spouse allowing the hero or heroine to marry his or her real love is also a sign of blameable authorial laziness.

The psychology of readers is not, however, finely attuned to probability theory. Suppose the author has the choice between getting from A to B in a plot in two steps or in six steps. For specificity, suppose that the two steps require events that will occur with likelihood 0.9 and 0.2 respectively, whereas each of the six events will occur with likelihood 0.75. Assuming the events in each sequence to be independent of each other, the two-step sequence is more likely to occur (0.18 versus 0.178), yet only the six-step sequence will be seen as having the desirable downhill property. As Daniel Kahneman and Amos Tversky write, "the plausibility of a scenario depends much more on the plausibility of its weakest links than on the number of links."[27] I believe authors should respect this particular quirk of the readers, since it prevents them from resorting to facile but unlikely coincidences.

Even a downhill stream may have many twists and turns before it somewhere winds safe to sea. If it did not, following its course would not provide much of an experience. The author is obliged, therefore, to provide the necessary surprises for readers, and obstacles for the characters, to keep audience interest alive. The repertoire of stratagems is huge, too huge to be surveyed or even to be classified. Some of them are closely linked up with the genre. Within the theatre, comedy, drama and tragedy have different means at their disposal. Whereas comedy often relies on misunderstandings to generate tensions, drama and even more so tragedy may rely on ignorance. As misunderstandings are dissipated, felicity ensues; as ignorance is lifted, disaster occurs. Novelists can add their own voices to those of the characters to generate uncertainty, as long as they do not deliberately mislead the readers.

I am now in a position to explain what I mean by the "right interpretation" of a text. As I stated at the outset, this is a question of explanation. Since all explanations are causal (including those that cite intentions as causes) and since a cause must precede its effect, it follows that actual audience perceptions of the work are strictly irrelevant. Intended perceptions, by contrast, can be part of the explanation. Among the antecedent causes of the work, the authorial intention is not all that matters. Unconscious attitudes of the author may also influence the work. Thus

[26] Stendhal said he would be appreciated in a hundred years, and he was. Thus the rationality of creation may not be verified until much later.

[27] Kahneman and Tversky (1982, 207).

Jules Verne's *L'île mystérieuse* may have been shaped by his anti-racist intentions as well as by his racist prejudices.[28] For the sake of brevity, however, I shall limit myself to conscious intentions.

An interpretation of a work of literature, then, is a claim that important features of the work can be traced back to decisions that the author made for the purpose of enhancing the aesthetic value of the experience that some specific audience could be expected to derive from the work.[29] To make a claim of this kind, literary critics must proceed just like other scholars. They can appeal to drafts, when they exist, and to statements by the author about the work, Stendhal's marginalia being an example. They can appeal to other works by the same author, to see if a similar pattern of choices is observed. They can refer to contemporary works, to distinguish the conventions that frame choices from the choices themselves. They can draw on other contemporary sources to determine the audience expectations that may have constrained the author.

In doing all this, their method is in no way different from that of other historians. Like other historians, they face the problem that the data are essentially finite, the past not being amenable to experiments. And like them, they can try to minimize the temptations of data mining and overinterpretation by triangulating old sources, looking for new sources and drawing out novel implications of their interpretation to be tested against evidence. They may differ from other historians in that their interpretation more often, although not invariably, goes together with value judgments. Did the author succeed, or come closer to succeeding than to failing, in his or her aim of creating a local maximum of aesthetic value? Some writers, to be sure, do not have this aim. They may only be concerned with making money or propaganda, goals which have different rationality requirements. But if one can make out a plausible case for the hypothesis that the author had mainly aesthetic pretensions, it makes sense to ask, as with any other aim, how well they were realized.

Earlier, I said that authorial failures might be intelligible. Authors, I have argued, 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. We may blame them if they sacrifice the latter goal to the former—i.e., if they sacrifice causality to teleology—but we can still understand why they do so. Even if one were to argue that Hamlet's procrastination is causally implausible, it could still be made to seem teleologically intelligible in the light of Shakespeare's need to delay the vengeance until the end of the play. This, too, would be a piece of interpretation. Although obviously very different from an interpretation of the delay in terms of Hamlet's psychology and circumstances, it does answer the same question: why the delay? Although in a good work of literature everything can be explained twice over, imperfect works may only allow for one interpretation.

Let me conclude by citing an example of how interpretation may violate or ignore the demands of explanation. In Jenny Davidson's recent interpretation of *Mansfield Park*, she makes the provocative claim that Fanny Price has "much in common" with Lucy Steele from *Sense and Sensibility*.[30] Although many readers have found Fanny Price insufferably priggish and obsequious, she would appear on a naive reading (the one I would defend) to have nothing in common with the selfish and scheming Lucy Steele. Davidson, however, views Fanny's modesty as instrumental,[31] and claims that her "ability to conceal her thoughts turns out to be a highly effec-

[28] Carroll (1993).

[29] The general approach I take in this chapter is often accused of embodying an "intentional fallacy." Without going further into the matter let me only state that I agree with the responses of Noel Carroll (1992, 1997).

[30] Davidson (2004, 155, 164).

[31] Ibid., 150.

tive stratagem"[32] in the conquest of Edward Bertram. These claims fail two tests of intentionality. First, there is no evidence in the novel for imputing scheming intentions to Fanny Price. Although her modesty is in fact rewarded, that consequence of her behavior cannot explain it.[33] Second, there is no evidence for imputing to Jane Austen an intention to make readers view Fanny Price as similar to Lucy Steele. Davidson cites the facts that "Fanny" evokes Cleland's novel *Fanny Hill* (a "woman of pleasure") and that "Price" has obvious monetary connotations.[34] Although the text may cause these associations to be produced in some modern readers, Davidson offers no evidence that Austen intended her readers to associate Fanny Price with the heroine of a pornographic novel.[35] Her reading amounts in fact to a functional explanation of the text.

RATIONALITY AND OPACITY

In literary interpretation the assumption of authorial rationality, as shown in the interplay of final and efficient causes, can enhance the transparency of the work. When we try to interpret the behavior of historical persons, the assumption of rationality can make their actions harder to understand. Ultimately, the problem is not insurmountable, but it has to be faced.

It is well and good to claim that behavior must be explained in terms of the antecedent mental states—desires and beliefs—that cause them, but how do we establish these prior causes? On pains of circularity, we cannot use the behavior itself as evidence. We must look at other evidence, such as statements by the agent about his or her motivation, the consistency of his or her non-verbal behavior with these statements, the motives imputed to him or her by others, and the consistency of their non-verbal behavior with these imputations. Yet how can we exclude the possibility that these verbal and non-verbal forms of behavior were purposefully chosen to make an audience believe, falsely, that a particular motivation was at work? Professions and allegations of motivations can themselves be motivated.

The rationality of the agents, in other words, can be an obstacle to imputing motives to them. This is turning a standard argument on its head. Donald Davidson has argued, notably, that we cannot impute motives or beliefs to an agent unless we assume that he or she is by and large rational. The discrepancy may be explained by Davidson's explicit stipulation that the agent to be interpreted states his motives and beliefs "honestly."[36] When trying to spell out the conditions for interpreting behavior, he assumed that the agents are not trying to deceive the interpreter. Strictly speaking, of course, social actors are usually not trying to fool the historian or the social scientist, but other actors to whom they stand in conflictual or potentially conflictual relations. Yet the historian, in particular, may not have much evidence beyond words or acts that may, for all he or she knows, have been produced for the purpose of inducing false beliefs about

---

[32] Ibid., 161.

[33] Also, the hypothesis of a mercenary Fanny Price cannot account for her rejection of a marriage proposal from the better situated Henry Crawford. Just imagine how Lucy Steele would have responded!

[34] Davidson (2004, 163–64).

[35] For a more sustained effort along the same lines, see Heydt-Stevenson (2000). Her argument amounts to a claim of an almost systematic attempt by Austen to deceive naive readers of *Mansfield Park* into believing in the innocence of Fanny Price, while allowing more perspicacious readers to understand that she is "little more [*sic*] than a fetishistic commodity, essentially [*sic*] bought and sold by members of her family, encouraged to prostitute herself for rank and wealth" (328). Like Davidson she ignores the inconvenient fact of Henry Crawford's rejected offer, and like her she cites the use of "Fanny" and "Price" as evidence for the sex-for-money interpretation of Fanny Price.

[36] Davidson (1980, 290).

motives in contemporaries. Often, the situation will be that of a "pooling equilibrium" or "cheap-talk equilibrium" in which sincere and insincere professions coincide.[37]

Before I proceed it may be useful to distinguish two dimensions of sincere motivations: strength and depth. Strength is measured by the sacrifices an agent is willing to make of other interests in order to satisfy the motive in question, and depth by the stability of the motive over time. The two often go together, but need not. The fanatical Communist may become an equally fanatical anti-Communist, being willing to sacrifice a great deal in either situation. Conversely, a person may be a steady churchgoer but donate little when the collection box comes around. The extreme of insincere motivation is represented by the person who is willing to profess instant loyalty to any cause if it serves his or her interest but is not ready to sacrifice anything to it.

As my main example I shall consider the motivations of participants in civil wars. Disregarding the (often important) motivation of revenge, I shall consider the motives of money, religion and politics. These are not mutually exclusive. One may seek political power to impose one's religion, or seek money to fund activities directed towards one of the other ends. Even conceptually, the lines can be blurred. Referring to the British civil war Austin Woolrych writes that "the idea of politics and religion as alternative grounds would have seemed strange to most seventeenth-century minds, which saw the two as intertwined."[38] At least, this is what they would say about their opponents. In Britain, and even more strongly in France the previous century, the monarch claimed that heresy was treason. Conversely, just as British Puritans claimed that "absolutist tendencies in the state were part and parcel of popery,"[39] French Protestants had perceived "a royal plot against the liberties in general, including religious freedom."[40] These causal and conceptual links do not, however, prevent us from asking whether, in a given case, a professed motivation might be spurious or an alleged motivation nonexistent.

During the French wars of religion (1562–1598), the warring parties constantly accused each other of using religion as a pretext for their political or even pecuniary aims. Neutral observers such as Montaigne found hypocrisy and opportunism on all sides.[41] Historians have debated the matter of the sincerity of religious conviction ever since. Tocqueville claimed "most of the elite precipitated themselves into a change of religion out of calculated ambition or greed, whereas the common people adopted it from conviction and without any prospect of gain."[42] That may indeed be roughly true, but there are exceptions to both claims. Among the rank and file iconoclasts, some were motivated by greed and resentment. As in all civil wars, conflicts initiated at the central level easily became the pretext for settling accounts at the local level.[43] Conversely, there is little question that some members of the elite were committed to one or the other religious cause. Among the main Protestant leaders, the Prince de Condé was willing to sacrifice peace as well as national unity for the sake of the triumph of the faith, whereas for his

[37] For these ideas, see for instance Kreps and Sobel (1994).

[38] Woolrych (2002, 248–49).

[39] Ibid., 249.

[40] Jouanna (1998, 178).

[41] Montaigne (1991, 495); see also the passage cited in Babelon (1982, 410–11).

[42] Tocqueville (1955, 187; translation modified).

[43] Babelon (1982, 189); Constant (2002, 123–24); more generally Kalyvas (2006).

cousin Henri de Navarre (the later Henri IV) the religious goal ranked only third in importance.[44] I shall have more to say about Henri.

It is easy to cite instances that throw some doubt on the strength as well as the depth of religious conviction of leaders on both sides. Henri IV converted six times in his life, although only the last conversion, in 1593, could be (and was) suspected of opportunism. His father, Antoine de Bourbon, had already made it clear that his faith was for sale to the highest bidder. He accompanied the Queen Regent to mass, and his Protestant wife to communion. On his deathbed, he sought consolation from both religions. A leading reformer, Cardinal de Châtillon, married after his conversion but retained both his title as cardinal and the revenue from his bishopric. Another prelate, Antoine Carraciolo, bishop of Troyes, also wanted to combine a Protestant ministry with the income from his bishopric. A leading Catholic, Henri Duc de Guise, was perfectly willing to seek an alliance with the Calvinists against King Henri III.[45]

A peculiarly complex situation arose just before the first civil war erupted. As a last-ditch—in fact premature—attempt to achieve civil peace the Queen Regent of Medici called a religious colloquium at Poissy in 1561 to see if a theological compromise could be achieved.[46] The main issue concerned the presence of Jesus Christ in the bread and wine served at mass or communion. The most literal interpretation was the Catholic one, the dogma of transubstantiation; less literal was the Lutheran doctrine of consubstantiation; next, the Calvinist view that the presence, while "real," was only "spiritual" and not "corporeal"; and finally the position of Zwingli that the presence was merely "symbolic."

It may seem hard to understand today why this question was so important that people were willing to kill and be killed over it. A brief answer is that both the denial of transubstantiation and iconoclasm (see below) were part of a complex of beliefs that followed from the rejection of God's immanence in the world. Another part of this complex was the belief in predestination which, paradoxically, enabled Calvin's followers to acquire the subjective certainty of salvation.[47] By contrast, those who believed that salvation had to be earned lived in a state of constant anxiety because they could never be sure they had done enough. Calvin wrote in 1539, five years after his conversion, that even when he had satisfied the demands of the church to confess his sins and efface God's memory of them by doing good works and penance, "I was far removed from certainty and tranquility of conscience. For each time that I delved into myself or lifted my heart up to You, I was struck by such an extreme horror that neither purgations nor disculpations could relieve me."[48]

The main interlocutors at Poissy were Calvin's second in command Théodore de Bèze and the Cardinal de Lorraine, brother of the Duc de Guise who was the leading military commander on the Catholic side. They were flanked on each side by extremists. Lorraine had notably to contend with Diego Lainez, general of the Society of Jesus and "uncompromisingly committed to Rome."[49] Bèze, on his side, was constrained by the presence of the equally uncompromising Pe-

---

[44] Babelon (1982, 221).

[45] Examples in this paragraph from Babelon (1982) and Jouanna (1998).

[46] Nugent (1974) is the standard work.

[47] Weber (1958, 115).

[48] Cited after Jouanna (1998, 335).

[49] Nugent (1974, 120).

ter Martyr Vermigli, whom d'Espence, one of the most moderate Catholics, viewed as "a pure Zwinglian."[50]

During the discussions, Lorraine suggested several times that Bèze might be willing to accept the Augsburg confession, i.e., consubstantiation. The first time, Bèze responded by asking whether the Cardinal would subscribe to it first, to which the Cardinal made an evasive reply ("une response fort double").[51] The next exchange was more substantial, but no less inconclusive:

> [Lorraine] returned to the topic of the Augsburg confession, asking the [Protestant] ministers why they were not willing to accept it. They replied that it was not reasonable to request this of them, since he and others on his side did not approve it; but if they would subscribe to it first, it should easily be possible to reach an agreement. All the more so since they did not know whether [the confession] was offered them on behalf of all [the Catholic prelates] or on behalf of a single person [the Cardinal]. The Cardinal replied "I am not bound [*astraint*] to swear by the word of any master, which is why I subscribe neither to those who have made the Augsburg confession nor to you, while nevertheless being ready to subscribe both to them and to you if you speak what is true. Moreover, my friends who are here present can testify that I have not said anything to you but what is their common opinion," whereupon the latter, the cardinal having looked at them from one end to the other, showed no sign of agreement nor of disagreement. "Hence," said Bèze, "since you are not willing to subscribe to this confession, it is not reasonable to ask us to do so."[52]

In these exchanges "Bèze and Lorraine desperately tossed the formula back and forth in an effort to pin the blame for the impasse on the other side."[53] The assertion—implied by the statement "we'll do it if you do it first"—that they were willing to accept a compromise position cannot be taken as proof of sincere willingness, as each party knew that in the presence of the hardliners the other was not going to accept it. Although their opponents in the debates were unlikely to be deceived by the posturing, the public at large might well be. Hence this was a purely strategic use of argument,[54] with little evidentiary value. A similar pseudo-exchange of views occurred in a colloquium in 1562 in Saint-Germain, where the topic was religious practice rather than dogma. Although the Calvinists appeared to propose a compromise solution—religious images were to be allowed on the outside of churches, but not inside—they probably did so only because they knew it would not be accepted. Soon after, the failure of the second colloquium destruction of images occurred all over France, triggering the first of the eight civil wars.

With a slight exaggeration one might say that the French wars of religion began and ended with "arguing and bargaining over transubstantiation." The failed colloquium of 1561 was the beginning. The end came with the conversion of Henri IV to Catholicism in 1593.[55] Although Henri never said that "Paris is worth a mass," it seems clear that he converted, in part, for instrumental reasons. His adherence to the Christian religion in general, regardless of doctrine, seems to have been deep but not strong. His commitment to Calvinism, from which he converted three times (in 1562, 1572, and 1593), is more of an open question. There is evidence,

---

[50] Ibid., 1446.

[51] Bèze (1882, 319); Nugent (1974).

[52] Bèze (1882, 325).

[53] Nugent (1974, 160).

[54] Elster (1995).

[55] See Feret (1875), Wolfe (1993), and Love (2001) for various aspects of this episode.

though, that at times he adopted the tendency of the Calvinists to see themselves as "the chosen instruments and special vessels of God" rather than, as did the Catholics, "in a passive light as the receivers of grace."[56]

Yet a Calvinist king on the French throne was a political impossibility. In the spring of 1593, Henri engaged in discussions with Catholic and Protestant theologians, "as if the colloquium of Poissy was being reborn from the ashes."[57] On July 23, the day of his abjuration from Calvinism, he met with four bishops and debated theology with them for five straight hours. He refused to accept the doctrine of the purgatory, and expressed reservations about the permanent "real presence" in the sacramental bread, outside the hours of church service. The refusal of the purgatory may have been needed to prevent the Huguenots from rising up in arms, since that doctrine was widely seen as a source of intolerable abuses in the Church. His reservations or hesitations about the real presence were related to questions about transubstantiation and the risk of alienating the Calvinist critics of that doctrine. Although it is impossible to determine how much he was influenced by the arguments of the theologians and how much by strategic considerations, a reasonable conjecture is that his strategic needs made him especially open to those arguments. Subjectively, he may have believed that his conversion was sincere.

I have cited these examples to bring out how professions (or denials) of motives can be made for instrumental or strategic reasons that detract from their value as evidence for mental states. I suspect that this difficulty may be at the root of the lack of concern for motivation among some social scientists. Instead of trying to establish these elusive antecedents of action directly, they impute them on the basis of the consequences of action. The result is a kind of rational-choice functionalism that has some similarity with the interpretations of Austen I discussed earlier. Let me use as an example the work of Gary Becker and Casey Mulligan on endogenous rates of time discounting.[58] They claim, plausibly, that if people attach high value to future consequences of present behavior, i.e., have a low rate of time discounting, their lives go better. They also assert, not implausibly, that higher education may shape time preferences in that direction. They conclude, implausibly, that people "may *choose* greater education in part because it tends to improve the appreciation of the future, and thereby reduces the discount of the future" (my emphasis). Yet they do not cite a single statement by a person asserting that he or she embarked on higher education in order to acquire a lower rate of time discounting. The "choice" they refer to is simply imputed to the agents on the basis of the consequences. Also, they ignore the conceptual problem that people who do not care about the future would not be motivated to take actions to make them care more about it.

The problems of imputing motives are not insurmountable, however.

As long as there is a fact of the matter—as long as these mental states really exist—it is in principle possible to discover it. In the case of Antoine de Bourbon, there may have been no fact of the matter. His beliefs were both so weak and so shallow that his religious behavior may have been little more than situation-triggered reflexes. In many cases, however, historians have been able to plausibly impute motives, using a wide array of techniques and procedures.

One technique is to go beyond statements made before an audience and to look for those less likely to be motivated by a desire for misrepresentation. Letters, diaries, reported conversations and observations by third parties can be invaluable correctives. These sources are used not

[56] Love (2001, 273–74).

[57] Babelon (1982, 557).

[58] Becker and Mulligan (1997).

only by those who study the actors in question, but also by those who may be affected by their behavior. The French kings, for instance, had a "black cabinet" whose function it was to intercept and read private letters, to discover whether those who expressed support for their policies in public might be undermining them in private. If we can tell that the reason why some delegates to the French Assemblée Constituante in 1789 voted against bicameralism and royal veto is that they feared for their lives if they voted otherwise, it is because we have access to letters they wrote to their wives. In the assembly, they appealed to the public interest. Some personal diaries of the delegates also have great value.

In trying to excavate the motivations behind the massacre on St. Bartholomew's night in 1582 historians have found it useful to go beyond the biased accounts of the participants and rely on reports by diplomats who had an interest in getting it right.[59]

Still another technique is to go beyond published documents to consult drafts that may contain more uncensored thoughts and fewer hypocritical ones. The published version of Marx's *The Civil War in France*, for instance, is much less critical of the Paris Commune than the drafts published long after his death. A similar relation obtains between his official letter to Vera Sassulitch and his draft letters. Government documents written for internal circulation are also likely to be more revealing than bland public statements, which is of course why governments often refuse to hand them over to investigators. Fortunately for historians, the tendency in Western democracies is for governments to refrain from opening the archives of their predecessors lest their successors do the same to them.

There is also a sharp contrast between what actors may say in public and what they say behind closed doors. Although the published debates of the Assemblée Constituante are endlessly fascinating, two factors conspire to make them less than reliable as evidence about mental states. On the one hand, the public setting constrained the delegates to use public interest arguments only; naked group interest was inadmissible. On the other hand, their vanity was stimulated by speaking before a thousand fellow delegates and a thousand auditors in the galleries. In both respects, the Federal Convention was more conducive to secrecy. Because the number of delegates was small (55, compared to 1200 in Paris) and the proceedings shrouded in secrecy, interest-based bargaining could and did occur. At the same time, as Madison wrote many years later, "Had the members committed themselves publicly at first, they would have afterwards supposed consistency required them to maintain their ground, whereas by secret discussion no man felt himself obliged to retain his opinions any longer than he was satisfied of their propriety and truth, and was open to the force of argument."[60] Nor did the fear of future revelations chill the debates, as the secrecy was supposed to extend indefinitely and was in fact broken only by the publication of Madison's notes many decades later. Strategic reasons for misrepresentations are blunted if sincerity carries no cost.

Finally, we can try to determine whether the non-verbal behavior of the agents is consistent with their professed motivation. Do they put their money where their mouth is? Some behavioral patterns may, for instance, reveal the true motivation of kidnappers who claim to act for ideological motives. In 1996 in Costa Rica, kidnappers (mainly ex-contras from Nicaragua) demanded a $1 million ransom in addition to job guarantees for workers, a cut in food prices, a rise in the Costa Rican minimum wage, and the release of fellow rebels from prison. When they were offered $200,000, "they were satisfied . . . and did not insist on the release of four convicted kid-

[59] Constant (2002, 99); see also the masterful account in Jouanna (2007).

[60] Farrand (1966, 3:479).

nappers from jail or the freeze of utility rates or the pay raise for government workers—a fact that persuaded the authorities that their Robin Hood/rebel stance was a ruse and that money had always been their goal."[61]

Another example is provided by the behavior of French aristocratic émigrés in London during the French Revolution. In this hotbed of rumor and competition to be more-royalist-than-thou, one had to convey one's confidence that the counterrevolution was imminent. Verbal assurances were not enough. "Any person who rented an apartment for more than a month was badly regarded; it was better to rent by the week to leave no doubt that one was ready to be called back to France by the counterrevolution."[62] Not only contemporaries, but also historians routinely use such behavioral indicators to judge the sincerity of public professions of loyalty. Towards the end of World War II, for instance, there was a marked degree of skepticism in occupied France about the prospects for German victory. It might not be safe to express this attitude, but it was reflected in behavior. The proportion of high school students who chose German as a foreign language (or whose parents chose it for them) doubled from 1939 to 1942, while falling rapidly thereafter.[63] Many publishers who eagerly signed up for the right to translate German books chose not to use the option.[64]

Judges and jurors often proceed in the same way. Sometimes, they ask, "Did the accused have a motive for doing X?" hoping that an answer will help them decide whether he in fact did X. In this case, "having a motive" is an objective idea, namely whether the accused would in some way benefit from doing X. In other cases, more relevant here, it is established that the accused did X and their question is "What was his motive for doing it?" To establish whether a killing was a crime of passion or a coldblooded action they do not mainly look at objective benefits, but try to establish the subjective state of mind of the accused. If the accused claims to have acted in a fit of anger or jealousy and later is shown to have bought the murder weapon in advance, his credibility is weakened.[65]

These techniques may fail if they have been anticipated and exploited by the agents. The existence of the "black cabinet" did not remain unknown for long. Persons whose mail might be subject to interception learned to be careful; moreover, they could exploit the system by writing damaging lies about their enemies in their correspondence. In the émigré example, both true believers and disbelievers would be motivated to lease by the week, the former to facilitate their return to France when the day came and the latter to escape criticism of being defeatist. Taking out a long lease can be a strategic move to make opponents believe one has a long time horizon rather than a naive expression of a patient attitude. In the Paris-based negotiations between the US and North Vietnam, the Vietnamese made a good opening move when they took out a two-year lease on a house, thus signaling that they were not in any hurry. Henry IV knew that he had to take his time converting, as a too quick change of heart would appear suspect.[66]

There are limits, however, on people's ability to weave the tangled web of deceit without revealing their true motives. Hypocrisy, Somerset Maugham said, is a full-time profession. Even

[61] Auerbach (1998, 284).

[62] Boigne (1999, 1:142).

[63] Burin (1995, 304–5).

[64] Ibid., 333. Another behavioral indicator of confidence or the lack of it is provided by the stock exchange movements (Destrem and Destrem 2003, 90).

[65] Sass (1983) lists thirteen reasons why a claim to have committed a crime out of passion might lack credibility.

[66] Jouanna (1998, 382); see also the letter by Henri IV cited in Babelon (1982, 333).

Tartuffe slipped in the end. To argue for the sincerity of Henri IV's religious beliefs, Jean- Pierre Babelon not only quotes the positive evidence of "numerous episodes where his religious spirit manifested itself without any advertising intention (*intention publicitaire*)," but also argues "had there been any hypocrisy, it would have showed its horns on this or that pleasant occasion."[67] Along the same lines we may quote Montaigne:

> Those who counter what I profess by calling my frankness, my simplicity and my naturalness of manner mere artifice and cunning—prudence rather than goodness, purposive rather than natural, good sense rather than good hap—give me more honour than they take from me. They certainly make my cunning too cunning. If anyone of those men would follow me closely about and spy on me, I would declare him the winner if he does not admit that there is no teaching in his sect which would counterfeit my natural way of proceeding and keep up an appearance of such equable liberty along such tortuous paths, nor of maintaining so uncompromising a freedom of action along paths so diverse, and concede that all their striving and cleverness could never bring them to act the same.[68]

In other words, while the benefits of misrepresentation may be considerable, the costs can be prohibitive. To some extent, the instrumental profession of motives is self-limiting. Because any given motive is embedded in a vast network of other motives and beliefs, the number of adjustments to be made in sustaining hypocrisy can be crippling. A single false note may be enough for the whole construction to crumble.[69] For this reason, among others, Descartes may have been right: "the greatest subtlety of all is never to make use of subtlety."[70] A

ACKNOWLEDGMENTS

I am grateful to John Ferejohn, Raquel Fernandez, Joseph Frank, Dagfinn Føllesdal, Russell Hardin, Stephen Holmes, Steven Lukes, Adam Przeworski, John Roemer and Peter Stone for their comments on an earlier version of this paper.

REFERENCES

Aristotle. 1962 [1920]. *On the art of poetry*. Trans. Ingram Bywater, with a preface by Gilbert Murray. Oxford: Clarendon Press.

Auerbach, A. H. 1998. *Ransom: The untold story of international kidnapping*. New York: Henry Holt.

Babelon, J.-P. 1982. *Henri IV*. Paris: Fayard.

Becker, G., and C. Mulligan. 1997. The endogenous determination of time preferences. *Quarterly Journal of Economics* 112:729–58.

Bekhtereva, N. P., M. G. Starchenko, V. A. Klyucharev, V. A. Vorob'ev, S. V. Pakhomov, and S. Medvedev. 2000. Study of the brain: Organization of creativity. *Human Physiology* 26:516–22.

Bèze, T. de. 1882. *Histoire ecclésiastique des églises réformées*. Toulouse: Société des livres religieux.

Boigne, Comtesse de. 1999. *Mémoires*. Paris: Mercure de France.

Budd, M. 1995. *Values of art*. London: Allen Lane.

Burin, P. 1995. *France à l'heure allemande*. Paris: Seuil.

---

[67] Babelon (1982, 554).

[68] Montaigne (1991, 897).

[69] Mackie (1998).

[70] Descartes (1646, 636).

Cannon, C. 1971. "As in a theater": Hamlet in the light of Calvin's doctrine of predestination. *Studies in English Literature* 11:203–22.

Carroll, N. 1992. Art, intention and conversation. In *Intention and interpretation*, ed. G. Iseminger, 97–131. Philadelphia: Temple University Press.

———. 1993. Anglo-American aesthetics and contemporary criticism. *Journal of Aesthetics and Art Criticism* 51:245–52.

———. 1997. The intentional fallacy: Defending myself. *Journal of Aesthetics and Art Criticism* 55:305–10.

Constant, J.-M. 2002. *Les Français pendant les guerres de religion*. Paris: Hachette.

Davidson, D. 1980. *Essays on actions and events*. Oxford: Oxford University Press.

———. 1993. Locating literary language. In *Literary Theory after Davidson,* ed. R. Dasenbrock, 295–307. University Park: University of Pennsylvania Press.

———. 2004. *Problems of rationality*. Oxford: Oxford University Press.

Descartes, R. 1646. To Elisabeth, January 1646. In *Oeuvres Philosophiques de Descartes*, ed. F. Alquié. Paris: Classiques Garnier.

Destrem, P., and D. Destrem. 2003. *A la botte. La Bourse sous l'Occupation*. Lausanne: L'Age d'Homme.

Doris, J. 2002. *Lack of character*. Cambridge: Cambridge University Press.

Elster, J. 1975. *Leibniz et la formation de l'esprit capitaliste*. Paris: Aubier-Montaigne.

———. 1983. *Explaining technical change*. Cambridge: Cambridge University Press.

———. 1995. Strategic uses of argument. In *Barriers to conflict resolution*, ed. K. Arrow, R. Mnookin, L. Ross, A. Tversky, and R. Wilson, 236–57. New York: Norton.

———. 2000. *Ulysses unbound*. Cambridge: Cambridge University Press.

Empson, W. 1982. *Essays on Shakespeare*. Cambridge: Cambridge University Press.

Farrand, M., ed. 1966. *Records of the federal convention*. New Haven, CT: Yale University Press.

Feret, Abbé. 1875. *Henri IV et l'Eglise catholique*. Paris: Librairie Victor Palmé.

Fernandez, R. 1981. La méthode de Balzac. In *Messages*, 54–69. Paris: Grasset.

Føllesdal, D. 1979. Hermeneutics and the hypothetico–deductive method. *Dialectica* 33:319–33.

———. 1982. The status of rationality assumptions in interpretations and in the explanation of action. *Dialectica* 36:301–16.

Grice, P. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Heydt-Stevenson, J. 2000. "Slipping into the ha-ha": Bawdy humor and body politics in Jane Austen's novels. *Nineteenth-Century Literature* 55:309–39.

Jouanna, A. 1998. Le temps des guerres de religion en France (1559–1598). In *Histoire et Dictionnaire des Guerres de Religion*, ed. A. Jouanna, J. Boucher, D. Biloghi, and G. Le Thiec, 3–445. Paris: Laffont.

———. 2007. *La Saint-Barthélémy: Les mystères d'un crime d'Etat*. Paris: Gallimard.

Kahneman, D., and A. Tversky. 1982. The simulation heuristic. In *Judgment under uncertainty*, ed. D. Kahneman, P. Slovic, and A. Tversky, 201–9. Cambridge: Cambridge University Press.

Kalyvas, S. 2006. *The logic of violence in civil war*. Cambridge: Cambridge University Press.

Kreps, D., and J. Sobel. 1994. Signalling. In *Handbook of game theory*, ed. R. Aumann and S. Hart, 2:849–68. Amsterdam: Elsevier.

Langfeldt, G., and Ø. Ødegård. 1978. *Den Rettspsykiatriske Erklæring om Knut Hamsun*. Oslo: Gyldendal.

Leibniz, G. W. 1969. Considerations on vital principles and plastic natures. In *Philosophical papers and letters*, ed. L.E. Loemker, 586–91. Dordrecht: Reidel.

Love, R. 2001. *Blood and religion: The conscience of Henri IV*. Montreal: McGill- Queen's University Press.

Mackie, G. 1998. Are all men liars? In *Deliberate democracy*, ed. J. Elster, 69–96. Cambridge: Cambridge University Press.

Mischel, W. 2004. Towards an integrative science of the person. *Annual Review of Psychology* 55:1–22.

Montaigne, M. de. 1991. *Essays*. Harmondsworth: Allen Lane.

Nugent, D. 1974. *Ecumenism in the age of the Reformation: The colloque of Poissy*. Cambridge, MA: Harvard University Press.

Parsall, P. 2001. Unfinished business: The problem of resolution in printmaking. In *The unfinished print*, ed. P. Parshall, S. Sell, and J. Brodie, 9–54. Washington, DC: National Gallery of Art.

Proust, M. 2003. *Finding time again*. London: Penguin.

Rimmon-Kenan, S. 1983. *Narrative fiction*. London: Methuen.

Ross, L., and R. Nisbett. 1991. *The person and the situation*. Philadelphia: Temple University Press.

Sass, H. 1983. Affektdelikte. *Nervenarzt* 54:557–72.

Stendhal. 1952. *Romans et Nouvelles*. Vol. 1. Paris: Gallimard.

Tocqueville, A. de. 1955. *The old regime and the revolution*. New York: Anchor Books.

Wagenknecht, E. 1949. The perfect revenge—Hamlet's delay. A reconsideration. *College English* 10:188–95.

Weber, M. 1958. *The Protestant ethic and the spirit of capitalism*. New York: Scribner.

———. 1969. The interpretive understanding of social action. In *Readings in the philosophy of social sciences*, ed. M. Brodbeck, 19–33. London: Macmillan.

Wolfe, M. 1993. *The conversion of Henri IV*. Cambridge, MA: Harvard University Press.

Woolrych, A. 2002. *Britain in revolution, 1625–1660*. Oxford: Oxford University Press.