3

Emotions and Rationality

Jon Elster

ABSTRACT

The issue of emotion versus rationality is intertwined with the more traditional issue of passion versus reason. This chapter discusses the impact of emotion on impartial reasoning and then considers how instrumental rationality can be affected by emotion, drawing on ancient and modern moralists as well as on recent work on emotion and decision making.

REASON AND RATIONALITY

The relation between emotion and rationality is a tangled one. To untangle it, we first need to distinguish between *reason* and *rationality*. From antiquity onward, philosophers and moralists have opposed reason and passion. By "passion," they mean more or less what modern writers mean by "emotion." What they mean by "reason" varies across writers and periods. The ancient moralists, of whom I shall take Seneca as my example, had a more restricted idea of reason than what we find, for instance, in James Madison. Both conceptions of reason differ, however, from the modern conception of rationality. The relation between emotion and reason or rationality naturally varies with the way we understand the latter.

Given his vast influence on later writers, we may take Seneca as an exponent of the classical opposition between reason and passion. His treatise *On Anger*, in particular, has a number of instructive observations.[1] He was well aware, for instance, of the distinction between occurrent emotions and emotional dispositions: "The difference between [anger] and irascibility is evident; it is like the difference between a drunken man and a drunkard,

[1] See Harris (2001) for the broader context of this treatise.

I am grateful to Kent Berridge and Olav Gjelsvik for their comments on an earlier draft of this chapter.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

between a frightened man and a coward. An angry man may not be an irascible man; an irascible man may, at times, not be an angry man" (I.iv). He also insisted on the involuntary and often all-consuming nature of the emotions: "they await no man's gesture and are not possessed, but possess" (I.xvi).

Seneca further observed that "Reason wishes the decision that it gives to be just; anger wishes to have the decision which it has given seem the just decision" (I.xviii). Here, Seneca is not only making a distinction between reason and passion as motivations; he is also claiming that, even from the point of view of the passionate agent, reason is the superior motivation. When la Rochefoucauld said that "Hypocrisy is a tribute vice pays to virtue" (Maxim 218), he did not go that far. He implied only that the vicious man must present himself to *others* as being motivated by reason. Seneca is claiming that he must also present himself to *himself* as being swayed by this motivation. There is a normative hierarchy of motivations that does not merely regulate our behavior in public but our self-image as well.

Seneca actually *defined* passion in terms of its capacity to overcome reason: "A man thinks himself injured, wishes to take vengeance, but dissuaded by some consideration immediately calms down. This I do not call anger, this prompting of the mind [*motum animi*] which is submissive to reason; anger is that which overleaps reason and sweeps it away" (II.iii). The spontaneous urge of the angry man to take revenge, or of the envious man to destroy the object of his envy, is a brute fact of human nature – neither rational nor irrational.[2] These urges are not passions. They become passions only when they override reason rather than submitting to it. Yet even when overriding reason, passion wants, as we have seen, to justify itself by reason.

Seneca also made one of the first theoretical comments on precommitment or self-binding:

> While we are sane, while we are ourselves, let us ask help against an evil that is powerful and oft indulged by us. Those who cannot carry their wine discreetly and fear that they will be rash and insolent in their cups, instruct their friends to remove them from the feast; those who have learned that they are unreasonable when they are sick, give orders that in times of illness they are not to be obeyed. (III.xiii)

Reason, in other words, can take precautions against passion. The converse cannot happen, for two reasons. First, as noted, passion stands in a relation of deference to reason. Second, anger is typically too shortsighted or

[2] Observations to this effect by Aquinas, Dr. Johnson, and Kant are cited in Elster, 1999a, p. 168.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

myopic to engage in strategic interaction with reason.[3] Reason, by contrast, is farsighted, at least in the modern conception.

As Seneca did not propose an explicit definition of reason, let me try to reconstruct one. In the reconstruction, the ancient idea of reason is defined by four features: orientation toward the future, lack of urgency, accurate belief formation, and justice or impartiality. (In the modern idea of reason, the first feature is modified.) The reasonable person is swayed by the future rather than by the past. Referring to someone who is free from anger, Seneca writes that

> in every case of punishment he will keep before him the knowledge that one form is designed to make the wicked better, the other to remove them; in either case he will look to the future, not to the past. For as Plato says, "A sensible person does not punish a man because he has sinned, but in order to keep him from sin; for while the past cannot be recalled, the future may be forestalled." (I.xix)

Before I define urgency, let me define a more general category of *prudence*, which I shall understand as the absence of both *impatience* and *urgency*. I define impatience as a preference for early reward over delayed reward; hence the prudent person is not only oriented toward the future, but toward the distant future, in the sense of being able to be swayed by long-term consequences of present behavior. I define urgency as a preference for earlier action over delayed action. Thus understood, impatience and urgency are independent phenomena: each can occur without the other. Seneca denounces urgency in many places, but there is no unambiguous denunciation of impatience.[4] I do not know the corpus of classical writers well enough to assert whether this imbalance is characteristic. In modern writers we certainly find the opposite imbalance: impatience is routinely denounced, urgency more rarely. Although proverbial wisdom recognizes the fact of "Marry in haste, repent at leisure," modern political philosophers and social scientists have paid less attention to the phenomenon.

A paradigmatic case of prudence as absence of urgency is the Roman general Fabius, the "hesitater," about whom Seneca asks:

> How else did Fabius restore the broken forces of the state but by knowing how to loiter, to put off, and to wait – things of which angry men know nothing? The state, which was standing then in the utmost extremity, had surely perished if Fabius had ventured to do all that anger prompted. But he took into consideration the

[3] I consider some exceptions to this statement in Elster, 2000, pp. 23–24.

[4] In an ambiguous denunciation, Seneca writes about Gaius Caesar that "so impatient was he of postponing his pleasure – a pleasure so great that his cruelty demanded it without delay – that he decapitated some of his victims by lamplight, as he was strolling with some ladies and senators on the terrace of his mother's gardens, which runs between the colonnade and the bank of the river. But what was the pressing need?" (I.xviii).

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

well-being of the state, and, estimating its strength, of which nothing now could be lost without the loss of all, he buried all thought of resentment and revenge and was concerned only with expediency and the fitting opportunity; he conquered anger before he conquered Hannibal. (I.xi)

One reason why urgency is harmful is that it impedes rational belief formation. (I shall later discuss other reasons.) "Reason grants a hearing to both sides, then seeks to postpone action, even its own, in order that it may gain time to sift out the truth; but anger is precipitate" (I.xvii). Here, anger causes suboptimal investment in information. In addition, anger prevents us from drawing correct inferences from the information that we do possess. Although the relevant passages are mostly too rhetorical to permit accurate analysis, it seems that Seneca had in mind that anger tends both to cloud the mind (without biasing it) and to bias it (e.g., through self-deception or wishful thinking).

Finally, reason is closely related to justice and fairness. As Seneca says, reason wishes the decision that it gives to be just. The kind of justice he has in mind is clearly consequentialist: the reasonable person "will openly kill those whom he wishes to have serve as examples of the wickedness that is slow to yield, not so much that they themselves may be destroyed as that they may deter others from destruction" (I.xix). To my knowledge, the idea that individual rights and duties might constrain or preempt consequentialist reasoning was unknown to him. For Seneca, the public interest was the dominant concern.

The next step in the development of the idea of rationality was taken, I believe, by the seventeenth-century French moralists. A particularly striking formulation can be taken from La Bruyère: "Nothing is easier for passion than to overcome reason; its greatest triumph is to conquer interest" (*Characters*, 4:77). In this maxim, La Bruyère distinguishes among three motivations: reason, passion, and interest. Although he does not tell us what he means by reason, we can perhaps reconstruct his meaning. When he says that reason is a weaker motivation than interest, in the sense of being more easily overridden by passion, it seems plausible to conclude that reason must be a disinterested or impartial motivation. We might ask a further question about the relative strength of these motivational forces. If passion can override both reason and interest, and with less resistance from reason than from interest, is interest also able to override reason? It does not follow, and La Bruyère does not tell us.

Writing a hundred years later, James Madison answered this question in the following terms:

To those who do not view the question through the medium of passion or of interest, the desire of the commercial States to collect, in any form, an indirect revenue from their uncommercial neighbors, must appear not less impolitic than it is unfair; since it would stimulate the injured party, by resentment as well as interest, to resort

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

to less convenient channels for their foreign trade. But the mild voice of reason, pleading the cause of an enlarged and permanent interest, is but too often drowned, before public bodies as well as individuals, by the clamors of an impatient avidity for immediate and immoderate gain.[5]

Although the text is not as crystal clear as one would wish, Madison seems to be saying that passion and interest *when impatient* are both able to override reason. If asked whether patient or long-term interest is capable of overriding reason, he would probably have answered that the two are unlikely to conflict. From Montaigne onward there was a long-standing idea in moral philosophy that long-term self-interest and morality tend to generate the same behavior. "Even if I did not follow the right road for its rightness, I would still follow it because I have found from experience that, at the end of the day, it is usually the happiest one and the most useful."[6] To some extent, therefore, the reference to an "enlarged and permanent interest" is redundant. If interest is permanent, it is also enlarged.

The language of interest is not yet the language of instrumental rationality. We have to add two elements and subtract one. The two elements to be added are (1) the idea of deliberate choice of means and (2) the extension of rationality to cover beliefs as well as actions. The element to be subtracted is the idea of self-interest.

The idea of interest does not by itself include the idea of a choice among several possible means to realize a given end. The pursuit of interest is sometimes described in words that would also apply to the activity of pursuing of a hare simply by running after it. It is a mechanical and mindless activity. In reality, the intelligent pursuit of one's interest may require considerable sophistication. In fact, even the pursuit of a hare may require a great deal of thought. Rather than just running in the direction where the hare is at any given moment, one may do better by running in the direction where it will be at some later point in time. In doing so, one may initially have to run away from the hare rather than toward it: "reculer pour mieux sauter." Similarly, the pursuit of economic welfare may require a sacrifice of welfare now for the sake of greater welfare later. More generally, the efficient pursuit of economic interest requires the agent to choose among alternative means. The profit-maximizing entrepreneur, for instance, has to choose among input combinations according to the relative prices of the factors of production. Ricardo or Marx never understood this principle. In the history of Western thought, the idea of the rational economic agent

[5] *Federalist*, no. 42.

[6] Montaigne, 1991, p. 709. At the Federal Convention in Philadelphia, George Mason, Gouverneur Morris, and Roger Sherman all appealed to arguments of this general kind (Farrand, 1966, vol. 1, pp. 49, 531: vol. 2, p. 3. (See also n. 29 below.) A modern game-theoretic form of this argument is that, although a rational egoist would never cooperate in a one-shot Prisoner's Dilemma, he or she might do so in an indefinitely repeated interaction.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

first emerged in the writings of the marginalists in the 1870s, and received a formal codification in the writings of Max Weber.

The pursuit of interest can hardly be called rational if it is based on irrational beliefs. For a belief to be rational, we should not require it to be true, only to be *well-grounded* in the evidence available to the agent. Actually, as Seneca noted, we should require a bit more: the agent should invest an optimal amount of resources (time, money, energy, etc.) in acquiring evidence or information. If these two conditions are satisfied, we might say that the agent pursues "enlightened self-interest." In contemporary as well as in earlier usage, however, that phrase is ambiguous. We may take it to mean that the agent acts on *rational beliefs*, but we could also use it to express the idea of acting according to *long-term interest*. Earlier, I noted that, as far as we can judge from Seneca, the ability to be motivated by long-term reward was no part of the ancient idea of reason. As I shall explain more fully, it is not part of the modern conception of rationality either. It is, however, part of the modern conception of reason.

At the same time, there is no need to limit instrumental rationality to the sphere of self-interest. Although both defenders and critics of rational-choice theory sometimes assume that rational agents must be self-interested, that assumption cannot be justified. Instrumental efficiency is obviously valuable in the pursuit of *any* end, whether self-interested or not. Altruism as well as hatred can fall within the domain of instrumental rationality. An individual who wants to alleviate poverty will naturally want to choose the charity that will use his donation most efficiently. He will avoid charitable organizations with high overhead costs and those whose funds end up lining the pocket of dictators rather than helping the poor. Unlike anger, which typically clouds or biases the mind, hatred is consistent with clear-headed instrumental rationality. The Holocaust – the greatest act of hatred in history – was carried out with notorious efficiency.

The spheres of reason and rationality overlap, in the sense that neither is a subset of the other. Instrumental rationality can be harnessed to the pursuit of private as well as public interest. Conversely, the modern concept of reason includes noninstrumental components. At the Federal Convention, Madison denounced "the immediate interest which one party may find in disregarding the rights of another or the good of the whole." What Madison calls "the good of the whole" is what I have referred to as the public interest. The notion of individual rights cannot, however, be directly assimilated to the public interest. Generally speaking, the idea of the rational pursuit of interest – private or public – is *consequentialist*. The rational agent chooses the means that will have the best consequences with regard to the interest he or she is pursuing or promoting. Rights, by contrast, are valued on nonconsequentialist grounds. The rights to life, liberty, security, and property are intrinsically valuable, over and above whatever value they might have on consequentialist grounds.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

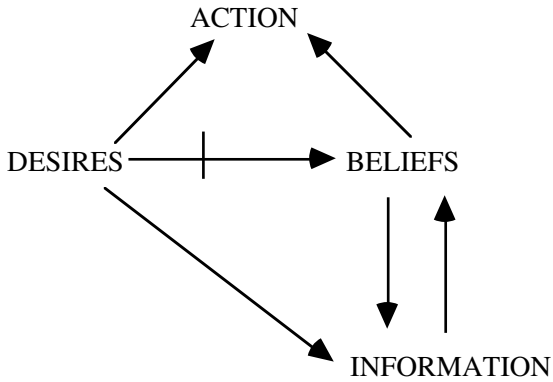


FIGURE 3.1. Rational-choice explanation of behavior

Rights and duties are correlative. The right to property, for instance, implies that other agents have a duty to abstain from taking it and that the state has a duty to prevent them from doing so. The right to work, if taken literally, implies that the state has a duty to provide employment for those who cannot find it in the labor market. Yet from Kant onward many have argued for duties that are not mere correlatives of rights, but independently justifiable. To the extent that these duties are prescribed universally and impartially, they, too, fall within the modern conception of reason. This conception, then, is richer than the ancient conception in two ways. It includes nonconsequentialist elements and it requires a long-term perspective. What has dropped out, it seems, is the idea that reason and urgency are incompatible.[7]

THE MODERN CONCEPTION OF RATIONALITY

We now are in a position to state more explicitly the modern theory of rational choice. The theory has two main aspects: it is explanatory and subjective. It aims at explaining behavior by assuming that agents do as well as they can by their own lights. The theory is summarized in Figure 3.1.

I have depicted a blocked arrow from desires to beliefs to indicate that processes such as wishful thinking and self-deception are inconsistent with rationality. Note, however, that the diagram leaves room for an indirect influence of desires on beliefs. How much evidence we collect depends on our prior beliefs about the expected costs and benefits of gathering new information. As indicated in the diagram, it also depends on our desires. To take a simple example, a person whose desires are very present-oriented,

[7] The idea, to be sure, is ignored, rather than denied.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

in the sense that he does not take much account of future consequences of present behavior, would not rationally invest many resources in finding out what those consequences might be.

This remark brings out the fact that rational-choice theory is radically subjective. One might want to say that someone who pays little attention to future consequences of present behavior is unwise, or downright stupid. His or her life is likely to be nasty, brutish, and short. Surely, we might want to say, drug addicts are not rational. Well, they can be. They are far from always rational, but the idea is not absurd. Imagine someone who for genetic or cultural reasons pays little attention to long-term reward and focuses almost totally on the short term. Under some circumstances, taking drugs might well be optimal. The person might even anticipate that drug taking will induce a deficit of rationality in the future, but decide that this effect is just a cost on a par with the financial and medical consequences of drug taking. All the conditions for rationality might well be fulfilled. The choice to become an addict could be a rational one. I do not believe this is true of most addicts, but it could be true of some.[8]

In light of what I have said, what kinds of things can be assessed as more or less rational? Actions, beliefs, and investment in information can be rational or not, optimal or suboptimal. But desires cannot be assessed as more or less rational. In the diagram, there are arrows going from desires, but no arrows going to desires; in the machinery of action, desires are the unmoved mover. This is not strictly true. To some extent we can choose our desires. Suppose I do not get any pleasure from classical music but observe that my friends do. I might then decide to expose myself to a great deal of classical music on the assumption that I will come to like it as much as they do. Although we should not say that the desire to listen to classical music is rational, the decision to develop that desire might well be. These cases are not common, but they exist.

There is a severe constraint, however, on the choice of desires. The consequences of having a new desire must be judged desirable in terms of the present desires, otherwise we would not want to develop the new desire. The classical music case satisfies that constraint. But here is a case that does not. Suppose I suffer from an inability to defer gratification, that is, from being unable to take account of future consequences of present behavior. And suppose scientists came up with a discounting pill, which would increase the weight of future rewards in present decisions. If I took the pill, my life would go better. In retrospect, I would be grateful I took the pill. But if I had a choice to take the pill or not, I would refuse. Any behavior that the pill would induce is already within my reach. I could stop smoking, start exercising, or start saving money right now, but I don't. Since I do not

[8] For a fuller discussion, see Elster, 1999b, chap. 5.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

want to do it, I would not want to take a pill that made me do it. Or, to put it differently, to want to be motivated by long-term consequences of present behavior *is* to be motivated by long-term consequences of present behavior.[9]

The person with a short time horizon is trapped. As the saying goes, "The eye cannot see beyond its horizon." A very myopic person may not be able to locate the optician who could provide him with the glasses he needs: he is trapped. In a similar way, a person can be caught in a belief trap. Gerry Mackie argues that

> women who practice infibulation [a form of female genital mutilation] are caught in a belief trap. The Bambara of Mali believe that the clitoris will kill a man if it comes in contact with the penis during intercourse. In Nigeria, some groups believe that if a baby's head touches the clitoris during delivery, the baby will die. I call these self-enforcing beliefs: a belief that cannot be revised, because *the believed costs of testing the belief are too high*.[10]

THE MODERN CONCEPTION OF EMOTION

I shall define occurrent emotions in terms of six features, which I shall first enumerate and then discuss at greater length.[11] They are:

- physiological arousal
- physiological expressions
- valence
- cognitive antecedent
- intentional object
- action tendency

When I discuss specific features, I draw attention to some of the ways in which they are relevant for the comparison with rationality. It is important to keep *all* the features in mind. Reducing emotions to "affect," that is, to arousal, can be misleading. Reducing them to their positive or negative "valence" can be equally misleading. The explanatory force of the emotions often lies in their fine grain. People behave quite differently under the influence of anger, indignation, resentment, hatred, or envy, although all these emotions are instances of negative affect.

By "arousal" I mean any departure from the physiological baseline – for instance, a decrease in the heart rate as well as an increase. This criterion is

[9] For a fuller development of this argument, see Elster, 2000, pp. 26–29, and Skog 2001.

[10] Mackie, 1996, p. 1009; italics added.

[11] This definition relies heavily on Nico Frijda's work, notably Frijda, 1986. In Elster 1999a, ch.II.2, I argue that Aristotle also imputed these six features to occurrent emotions. In ch.IV.2 of the same work, I elaborate on each of the six features and comment on various exceptions.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

important to distinguish emotions proper from other phenomena referred to by the same words. The term "fear," for instance, may denote visceral fear or fear proper, but it can also denote a simple belief-desire complex, as when I say I'm afraid it's going to rain.[12] Whereas visceral fear may short-circuit rationality, fear as a desire-belief complex will typically induce rational action, such as taking an umbrella. Similarly, the term "regret" may refer to a wrenching experience or to a mere wish that one had acted differently.

The fact that emotions have physiological expressions makes it possible for the emotion felt by one person to shape the behavior of other agents. When an expression is perceived as the forerunner of action directed toward another, the latter may take steps to defuse or encourage such action. If my antagonist can see from my face and posture that I am angry, he may back down. If a woman can infer from my facial expression that I may be about to make an advance to her, she may modify her behavior to encourage or discourage me. To the extent that these expressions are under the control of the conscious will, or can be approximated by deliberate simulation, they invite instrumentally rational exploitation.

The valence of the emotions – the pain and pleasure that accompany them – offers an obvious bridge to rational-choice theory. One might indeed think that the pain and pleasure of emotions are simply negative and positive utilities that may contribute to the overall expected utility of a given choice. For reasons I have spelled out elsewhere, I believe the picture is more complicated.[13] Here, let me simply note that the level of pain and pleasure derived from emotions are inversely related to the probability of the events that generate these emotions. As Barbara Mellers and her colleagues have shown, surprise is an important magnifier of affect.[14] In the choice between an action that is unlikely to generate high pleasure and one that is quite likely to do so, one might rationally choose the former.

The emotions I shall consider are indeed triggered by beliefs. The triggering beliefs may, like any other belief, be rational or irrational. Moreover, if the triggering belief is irrational, it may be because it is distorted by a prior emotion. When I compare your success with my failure, I experience a pang of envy. To rid myself of the feeling of inferiority, I tell myself a story according to which your success was due to immoral or illegal behavior. On the basis of this emotionally motivated belief I now feel indignation rather than envy. In such processes, emotions enter both as causes and as effects of beliefs.

[12] Gordon, 1987; see also Harris, 2001, p. 108.

[13] Elster, 1999a, pp. 154–156, 179–181, 301–306.

[14] Mellers, Schwartz, and Ritov, 1999. Paradoxically, perhaps, anticipation is also a magnifier of affect, as it enables us to savor the emotional experience before it actually occurs (Loewenstein, 1987).

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

TABLE 3.1. *Action Tendencies of Emotions*

Emotion	Action Tendency
Anger	Cause the object of anger to suffer (revenge)
Hatred	Cause the object of hatred to cease to exist
Contempt	Ostracism
Shame	"Sink through the floor"; suicide
Guilt	Confess; make repairs; hurt oneself
Envy	Destroy the envied object or its possessor
Fear	Flight; fight
Love	To approach and touch the other

Moreover, emotions have intentional objects: they are about something, or directed toward something. In the important case of emotions that target individuals, the object may be an action committed by a person or it may be that person's character. Anger, guilt, admiration, and pride are directed toward actions; hatred, contempt, shame, liking, and pridefulness are directed toward the person's character. The object of envy is not the envied person but the fact that he has something I want, in both senses of that word. But if what he has and I lack is moral goodness, or beauty, this distinction becomes tenuous.

Emotions, finally, have action tendencies. Some examples are given in Table 3.1.

Some emotions may be associated with wishes rather than with desires. One may wish that a certain state obtain, or one may desire to bring it about. In hatred, what matters is that the hated person or group disappear from the face of the earth. In envy, what matters is that the envied person lose his or her possessions. In neither case is there an additional emotional satisfaction derived by the state of affairs being realized through my agency. In hatred, I may take action if I am well placed to do so, but I might be just as happy, or even happier, if someone else did it for me. In envy, I may even have a positive preference for the other's ruin not coming about through my agency.[15] Some people who would abstain from calling the fire brigade if their neighbor's house were on fire would not themselves set fire to it (and not merely because of the risk of detection). By contrast, anger, contempt, love, guilt, and shame activate desires where it matters that they are realized through *my* agency. My desire for revenge is not slaked if the object is injured in a car accident. My desire for atonement is not alleviated if the person I have harmed wins the jackpot in a lottery. My feeling of contempt requires *me* to isolate myself from the object, not merely that the person be ostracized by others.

[15] Seneca, *On Anger* III.v.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

RATIONALITY AND EMOTION

Rationality, as summarized in Figure 3.1, requires three optimizing operations. First, the action must be the optimal means of realizing the agent's desire, given the beliefs. Second, the beliefs must be rational or well-grounded, given the evidence. Exactly what this means is controversial, but in many cases the general idea is unproblematic. Third, the agent must have invested optimally in acquiring the evidence that is needed to form a belief before acting. Violations of rationality may occur in any of these three optimizing operations. Such violations can have many causes. Here I only consider those which can be traced back to the interference of emotion with rationality. I have already mentioned many emotion-induced violations of reason that are also violations of rationality. These are emotionally induced cognitive flaws that arise through urgency, clouded thinking, or biased thinking.[16]

I now want to consider violations that are not due to flawed cognitive processes. To do so, I need to be more explicit about the objects of the "desires" or preferences that generate behavior. In the standard case, agents have primitive preferences over outcomes of action and outcome-induced preferences over actions. Who desires the end (and the side effects) desires the means. In other cases, agents have non-outcome-induced preferences over actions. In still other cases, agents have preferences over outcomes that are shaped by (their beliefs about) the actions that brought them about or, more accurately, that caused the set of outcomes to be restricted in a particular way.

In the standard case, assuming correct cognitive processing, the ensuing action will be rational. In making this assertion I ignore some famous literary cases of weakness of the will. Medea, when killing her children, says: "I know indeed what evil I intend to do. But stronger than all my afterthoughts is my fury."[17] Racine's Phèdre asks her confidante to "serve my fury, and not my reason." I do not know whether such states, characterized by the coexistence of lucidity and extreme passion, occur outside of literature.[18] I have more faith in Seneca's claim that "anger wishes to have the decision which it has given seem the just decision." Be this as it may, I shall focus on the two nonstandard cases.

Consider first non-outcome-induced preferences over actions. These are closely related to the action tendencies associated with emotion. Also, they

[16] Although we tend to think of emotional bias as taking the form of wishful thinking, the opposite bias is also possible. A French proverb says, "We believe easily what we hope *and what we fear*." Othello illustrates the latter, underexplored mechanism.

[17] See Harris, 2001, p. 169, n. 52, for some of the complications that arise in interpreting this text.

[18] Racine's *Andromaque* is unique among his plays in that it recognizes that passion tends to undermine the lucidity of the agent (see Elster, 1999a, pp. 111–117).

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

are closely related to the idea of urgency. In my earlier discussion, I have assumed that the impact of urgency is mediated by cognition, and in particular by the suboptimal investment in information gathering. Urgency can also have a direct impact, however. This is the case when the agent takes early action even when he or she knows (1) that the outcome of action would be better if the action were delayed; and (2) that the outcome itself will not be not delayed by delaying the action. By assumption (1), it is not a matter of faulty cognitive processing. By assumption (2), it is not a matter of impatience.

A more complicated and probably more common case is the following. The agent can act at time 1 or time 2. If action is taken at time 1, the outcome occurs at time 3. If action is taken at time 2, it occurs at time 4. (The times succeed each other in the numbered order.) The interval between time 1 and 2 is relatively long. The interval between time 3 and time 4 is relatively short. The outcome at time 4 is substantially better than the outcome at time 3. These vague phrases are intended to convey the idea that in a nonemotional state the agent would be able to defer gratification from time 3 to time 4. If the emotional state induces the agent to choose the earlier action and the smaller reward, the mechanism could be either urgency (a preference for an earlier action) or a temporary increase in the rate of time discounting (a stronger preference for an earlier reward).[19] If the choice of the earlier action is due to urgency, it violates the canons of rationality.[20] If it is due to increased impatience, it does not.[21] In practice, it might be hard to identify the culprit.

Social norms provide an important special case of non-outcome-induced preferences over actions. Consider two injunctions: "Always wear black in strong sunshine" and "Always wear black at a funeral." The former is instrumentally rational advice, as air circulates more quickly under dark

[19] In the related case of addiction, several writers have discussed the idea of endogenous changes in discounting rates (Becker, 1996, p. 120; Orphanides & Zervos, 1998; O'Donoghue & Rabin, 1999). The idea that emotions, too, can shorten the time horizon of the agents is more or less clearly stated by various writers. Yet the (crucial) question is rarely asked whether emotion undermines the agent's awareness of remote consequences or causes him to attach less weight to them in his utility function.

[20] A counterargument might be that someone who has an urgent desire to act will get rid of the urge by acting. Since having an unsatisfied urge is painful, and acting can relieve the pain, early action is rational if the gains from relieving the pain earlier rather than later offset the loss of receiving the smaller rather than the larger reward. The objection, in my view, is misguided. Although acting on the urge will relieve the urge, it is not typically instrumental action for the *purpose* of relieving the urge, in the way one can take Methadone to relieve oneself from a painful craving for heroin.

[21] Some might object to this statement. If a person who is normally able to defer gratification loses that ability in an emotional state, is that not the very paradigm of irrationality? There is indeed a temptation to assess desires and preferences as more or less rational on the basis of the way in which they have come about, but it is one we should resist. I am grateful to Kent Berridge and Olav Gjelsvik for pressing this point on me.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

clothes. The latter has no instrumental aspect at all. It is a social norm that is maintained by internalized emotions (shame) and sanctioning behavior by others in the community. Here are some other salient examples of social norms:

Norms regulating the use of money. (1) Suppose someone walks up to the person at the head of a bus queue and asks to buy that person's place. If the offer is accepted, two persons gain and nobody loses. If it is rejected, nobody loses. Yet there is a norm against making such offers. (2) In a small suburban community, Mr. H. mows his own lawn. His neighbor's son would mow it for $12. He wouldn't mow his neighbor's same-sized lawn for $20. This suggests that the value to him of the time it takes him to mow the lawn is both less than $12 and more than $20. The puzzle is resolved by noting the existence of a norm in such communities against monetary transactions of this kind among adults.[22]

Medical norms.[23] (1) Medical ethics tell doctors to treat the more severe cases first. Yet these are often (as in liver transplants) the patients who are too ill to benefit from treatment. (2) Doctors follow a norm of thoroughness, which tells them that once a patient has been admitted, he or she should get "the full treatment." Yet other patients might benefit much more from the time he spends on the later tests, since with respect to any given patient the doctor's time has decreasing marginal productivity. Both norms violate instrumental rationality with regard to the objective of saving lives or improving overall health.

The paradox of voting. Why do people bother to vote in national elections? Since no such election has ever been won by a single vote, an individual vote makes no difference to the outcome, and may entail considerable trouble for the voter. Yet people do vote in large numbers, to fulfill their "civic duty."

Norms of revenge.[24] People often seek revenge for an insult, even when (1) doing so is costly or risky, (2) they know that they will not be able to undo the harm that was done to them, and (3) they know the revenge will not dissuade others from harming them in the future. In other words, people seek revenge even when there are no tangible benefits and some tangible costs. Although many forms of revenge behavior occur spontaneously (this was Seneca's main concern), some societies have powerful social norms that reinforce the spontaneous tendency.

Norms of cooperation.[25] (1) The norm of fairness tells individuals to cooperate if and only if others do. The norm makes cooperation contingent on other people's behavior and not on the outcome of the agent's behavior. (2) The norm of "everyday Kantianism" tells individuals to cooperate if, and only if, it would be better for all if all cooperated than if nobody did. This norm makes cooperation contingent on hypothetical outcomes of behavior rather than on the actual outcome of the agent's behavior.

[22] Thaler (1980), from whom this example is taken, explains it in terms of loss-aversion. Yet as Amos Tversky pointed out to me a long time ago, the norm explanation is also plausible.
[23] For a fuller exposition, see Elster, 1992, pp. 91–94, 147–149.
[24] For a fuller exposition see Elster, 1999a, chap. 3, p. 4.
[25] For a fuller exposition, see Elster, 1989, chap. 3.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

I am making two claims about such norm-guided behavior. First, these norms violate instrumental rationality by subverting the goals of the agents who subscribe to them. Second, the norm following is induced or sustained by emotions. Both claims might be met by the objection that if people are sanctioned by others for violating the norm, norm following *is* instrumentally rational because it enables the agent to avoid punishment. I have two responses. First, we still have to explain why others sanction norm violators. The response that they do so because nonsanctioners will also be sanctioned (and so on at higher degrees) becomes implausible at more than two or three removes away from the original violation. Second, and more fundamentally, the objection rests on a misunderstanding of the impact of sanctioning. When I refuse to deal with a person who has violated a social norm, that person may suffer a financial loss. Far more important, however, the other will see the sanction as a vehicle for my emotions of contempt or disgust and suffer shame as a result. The material aspect of the sanction that matters is *how much it costs the sanctioner to penalize the target*, not how much it costs the target to be penalized. The more it costs me to refuse to deal with you, the stronger you will feel the contempt behind my refusal and the more acute will be your shame. The emotional meaning of sanctions was recognized by Aristotle, who wrote that "Shame is the imagination of disgrace, in which we shrink from the disgrace itself and *not from its consequences*" (*Rhetoric* 1384a; italics added).

In the second nonstandard case, agents have preferences over outcomes that depend on how they are brought about. I shall develop this idea by referring to the Ultimatum Game (Fig. 3.2).[26]

In this game, the first player proposes a division of ten dollars between himself and the second player. Proposals can be made only in whole dollars. The second player can reject the proposal, in which case neither player gets anything, or accept it, in which case the proposed division is implemented. In experiments, participants interact anonymously, through computer terminals. Also, in the experiments I shall consider there is only one-shot interaction, with no room for reputation building. Given these conditions, there is no reason to think that the players care about each other or that they might be concerned with building a reputation for the future. We would expect them to be concerned only with the payoffs in the experiment. If they are rational, and know each other to be rational, Player I will propose nine dollars for himself and one dollar for Player II, who will accept the proposal. What happens in experiments is quite different. Player I typically offers something like seven dollars for himself and three for Player II. Player II typically rejects any proposal that gives her less than three dollars.

Consider three interpretations of these findings, in terms of (a sense of) fairness, envy, and resentment. First, we could imagine that those in the

[26] For a survey of Ultimatum Game experiments, see Roth, 1995.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

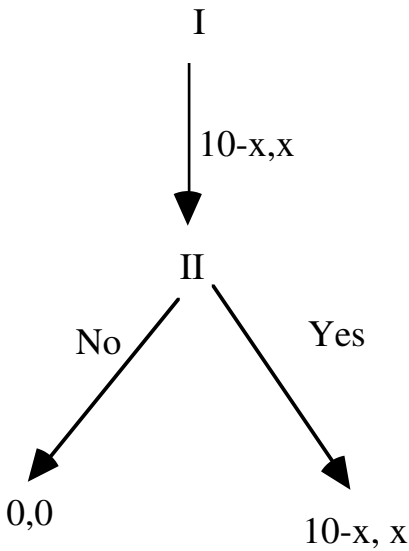


FIGURE 3.2. The Ultimatum Game

position of Player I are moved (although not only) by a sense of fairness when they abstain from making the most self-interested proposal, and that those in the position of II are moved (although not only) by fairness when they reject the most unfavorable proposals. Yet in experiments with the "Dictator Game," in which Player II simply has to accept the proposal without the option of rejecting it, Player I typically makes a more selfish proposal. This finding suggests that even in the Ultimatum Game Player I is motivated mainly by the anticipated rejection by Player II of an unfavorable proposal, rather than by a sense of fairness. Moreover, it is not really plausible to explain the rejection in terms of a sense of fairness. If an agent were genuinely moved by considerations of fairness, these would move him or her in the position of Player I as well as in that of Player II. Since participants are randomly assigned to the positions of Players I and II, there should not be any systematic difference in the sense of fairness displayed in their behavior. As those in the position of Player I are not (or not strongly) moved by fairness, it seems plausible to explain the behavior of those in the position of Player II in terms of envy or resentment rather than by a (genuine) sense of unfairness, and the behavior of those in the position of Player I by their anticipation of such envy or resentment.

Let us now consider the two emotion-based explanations in terms of envy and resentment. Whereas envy is an outcome-based emotion, resentment is action-based.[27] One can imagine that Player II, if facing an 8,2 proposal, might reject it out of envy. If that is the case, she should also

[27] Hirshleifer, 1987.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

reject the same proposal if Player I were constrained to choose, say, between (10,0) and (8,2). Intuition suggests and experiments confirm that Player II will be much more likely to reject in this constrained case than in the unconstrained case.[28] The explanation of the behavior in the Ultimatum Game must be resentment rather than envy.

To show the relevance of this issue, let me cite from a comment made by George Mason at the Federal Convention in Philadelphia. The discussion concerned whether the Western lands that would accede to the Union in the future should be admitted with the same rights as those of the original thirteen states. When Gouverneur Morris argued that they should be admitted as second-rate states, so that they would never be able to outvote the original states, Mason argued strongly for the opposite view:

> If the Western States are to be admitted into the Union, as they arise, they must be treated as equals, and subjected to no degrading discriminations. They will have the same pride & other passions which we have, and will either not unite with or will speedily revolt from the Union, if they are not in all respects placed on an equal footing with their brethren.[29]

Mason appeals to the "pride and passions" of the new states, not to their self-interest. Even if it would in fact be in their interest to accede to the Union on unequal terms rather than remain outside, they might still, out of resentment, prefer to stay outside. Implicitly, he appeals to the self-interest of the old states, not to their sense of justice.[30]

Is the resentful player in the Ultimatum Game who rejects the (8,2) proposal *irrational?* I cannot see any way in which the behavior can be said to serve his or her ends. An envious rejection might maximize a utility function in which the consumption of others enters on a par with the agent's own consumption, but I submit that there is no noncontrived utility function that would rationalize rejection out of resentment.

CONCLUSION

I have defended the traditional view that emotions interfere with and subvert instrumental rationality. This is not to say that we would be better off

[28] Camerer, 2003, chap. 3. 7. 5.

[29] Farrand, 1966, vol. 1, pp. 578–579. Sherman's veil-of-ignorance argument (n. 6 above) was made in the same context: "We are providing for our posterity, for our children & our grand Children, who would be as likely to be citizens of New Western States, as of the old States. On this consideration alone, we ought to make no such discrimination as was proposed by the motion."

[30] Actually, he used both arguments. By admitting the Western states on equal terms, the Framers would do "what we know to be right in itself" (Farrand, 1966, vol. 1, p. 578). To those who might not accept that argument, he added that the new states would in any case be unlikely to accept a degrading proposal.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

if we had no emotions. Emotions provide us with a sense of goal and direction in life, but also prevent us from going steadily in that direction. Except in unacceptably loose language, the first part of the previous sentence does not imply that "emotions are rational."

Could emotions *enhance* instrumental rationality? Although several writers have proposed arguments to this effect, I have not found any of them convincing.[31] The fact that efficient decision making and normal affect are impaired by the same brain lesions does not show that the latter is a condition for the former. The fact that an emotional action tendency can provide a more efficient response to danger than an exhaustive assessment of all options and their consequences shows only that the latter procedure is irrational, not that the former is superior to a rational response.

Might not emotions themselves fall under the scope of rational choice? I believe that occurrent emotions are largely unchosen. Although emotional dispositions are (at least partly) the result of action, I do not think we know enough about the causal mechanisms involved to be able to act efficiently on our dispositions. Even if we could, the costs might be prohibitive – five years of psychoanalysis can add up to twice the annual income of a middle-class professional.

More generally, there is in the social sciences, partly because of the dominating influence of economics, a pervasive tendency to "rationalize" all aspects of human behavior. This tendency is shown by the large number of "just-so" or "as-if" stories that, in my opinion, discredit their authors. We cannot deal rationally with irrationality if we refuse to accept its existence.

References

Becker, G. (1996). *Accounting for tastes*. Cambridge, MA: Harvard University Press.
Camerer, C. (2003). *Behavioral game theory*. Princeton, NJ: Princeton University Press.
Elster, J. (1989). *The cement of society*. New York: Cambridge University Press.
Elster, J. (1992). *Local justice*. New York: Russell Sage.
Elster, J. (1999a). *Alchemies of the mind*. New York: Cambridge University Press.
Elster, J. (1999b). *Strong feelings*. Cambridge, MA: MIT Press.
Elster, J. (2000). *Ulysses unbound*. New York: Cambridge University Press.
Farrand M. (Ed.) (1966). *Records of the Federal Convention*, vols. 1–3. New Haven, CT: Yale University Press.
Frijda, N. (1986). *The emotions*. Cambridge: Cambridge University Press.
Gordon, R. M. (1987). *The structure of emotions*. Cambridge: Cambridge University Press.
Harris, W. (2001). *Restraining rage: The ideology of anger control in classical antiquity*. Cambridge, MA: Harvard University Press.
Hirshleifer, H. (1987). The emotions as guarantors of threats and promises. In J. Dupré (Ed.), *The latest on the best* (pp. 307–326). Cambridge, MA: MIT Press.

[31] See comments on Damasio and de Sousa in Elster, 1999a, pp. 287–298.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. *Economic Journal*, *97*, 667–684.
Mackie, G. (1996). Ending footbinding and infibulation: A convention account. *American Sociological Review*, *61*, 999–1017.
Mellers, B. A.; Schwartz, A.; & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, *128*, 1–14.
Montaigne, M. de. (1991). *The complete essays* (Trans. M. A. Screech). Harmondsworth: Penguin.
O'Donoghue, T., & Rabin, M. (1999b). Addiction and self-control. In J. Elster (Ed.), *Addiction: Entries and exits* (pp. 169–206). New York: Russell Sage.
Orphanides, A., & Zervos, D. (1998). Myopia and addictive behavior. *Economic Journal*, *108*, 75–91.
Roth, A. (1995). Bargaining experiments. In J. H. Kagel and A. E. Roth (Eds.), *Handbook of experimental economics* (pp. 253–348). Princeton, NJ: Princeton University Press.
Skog, O.-J. (2001). Theorizing about patience formation. *Economics and Philosophy*, *17*, 207–219.
Thaler, R. (1980). Towards a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, *1*, 39–60.

https://doi.org/10.1017/CBO9780511806582.003 Published online by Cambridge University Press