In *The Brothers Karamazov*, Dmitry Karamazov tells the story of a lieutenant colonel who managed substantial sums on behalf of the army. Immediately after each periodic audit of his books, he took the available funds to the merchant Trifonov, who soon returned them with interest and a gift. In effect, both the lieutenant colonel and Trifonov benefited from funds that would otherwise have lain idle, producing no benefit for anyone. Because it was highly irregular, theirs was a secret exchange that depended wholly on personal trust not backed by the law of contracts. When the day came that the lieutenant colonel was to be abruptly replaced in his command, he asked Trifonov to return the last sum, 4,500 rubles, entrusted to him.

Trifonov replied, "I've never received any money from you, and couldn't possibly have received any" (Dostoyevsky 1982/1880, p. 129).¹

Dmitry Karamazov says that the lieutenant colonel implicitly trusted Trifonov. After his sad day of reckoning, the lieutenant colonel would presumably have said that Trifonov was not trustworthy. Unfortunately, Trifonov was trustworthy just so long as there was some longer-run incentive for him to be reliable in their mutually beneficial relationship. The moment there ceased to be any expectation of further gains from his relationship with the lieutenant colonel, Trifonov had no incentive to be trustworthy and, not surprisingly, he ceased to be. The lieutenant colonel might have thought Trifonov to be a man of honor who was, in that sense, trustworthy. If so, then, he clearly misjudged his man.

Had his dealings with Trifonov been reputable, the lieutenant colonel could have been protected by a contract that gave both parties
the right to be sued. “Who wants to be sued?” Thomas Schelling asks. Well, he notes in answer to his own question, “The right to be sued is the power to make a promise: to borrow money, to enter a contract, to do business with someone who might be damaged. If suit does arise, the ‘right’ seems a liability in retrospect; beforehand it was a prerequisite to doing business” (Schelling 1960, p. 43). The good right to be sued is the “power to accept a commitment.” It enables one to establish a strong commitment to fulfilling one’s half of a bargain. Trifonov had no right to be sued by the lieutenant colonel, who, indeed, could not even publicly accuse him. And therefore Trifonov could be assumed to have only a commitment to gain as much as he could from his dealings with the officer. That is all that the lieutenant colonel should have trusted Trifonov to do.

Writings on trust often take the view that it involves something beyond merely reasonable expectations based in self-interest. Many philosophers suppose we should distinguish our trust of another individual from our expectations about that individual’s behavior in particular respects (Baier 1986; Hertzig 1988). Many writers also suppose that trust is an inherently normative notion (Elster 1979, p. 146). We can make some sense of such claims by supposing that they are really misplaced claims about trustworthiness rather than about trust. You might be trustworthy in the very strong sense that you would reciprocate even when it was against your interest to do so, as Trifonov might have returned the final 4,500 rubles.

Even then, my trust of you must be grounded in expectations that are particular to you, not merely in generalized expectations. If I always trust everyone, then I do not meaningfully trust anyone. Trust is therefore in part inherently a rational or intentional commitment or judgment. My expectations about your behavior may be grounded in my belief in your morality or fairness or, most commonly, self-interest. With no prior knowledge of you, I may initially treat you as though I trust you, but our relationship can eventually be one of trust only if there are expectations that ground the trust. As Karamazov’s lieutenant colonel learned, inductively grounded expectations cannot be reliable for new contexts.

If the vague sense that trust requires more than rational expectations—expectations grounded in the plausible motivations of the trusted—is correct, then we are at a very early stage in the development of any theory to account for trust or even to characterize it. If the residual self-interest notion is largely correct, however, we already have the elements of a full-blown theory of trust that merely wants careful articulation and application. I will give an account of trust as essentially rational expectations about the—mostly—self-interested behavior of the trusted. The effort to construct such an account forces attention to varieties of interaction in which trust might arise and, hence, to differences in the plausible explanations of trust. The sense that trust requires more than reliance on the self-interest of the trusted may depend largely on particular kinds of interaction that, while interesting and even important, are not those of greatest interest in social theory.

I will present an account of trust as encapsulated interest, an account in which the trusted’s expectations of the trusted’s behavior depend on rational assessments of the trusted’s motivations. I will then argue that certain alternative, strongly-argued individual-level accounts of trust are implausible. Then I will turn to generalizations to trusting large numbers or institutions.

Trust as Encapsulated Interest

If Karamazov’s lieutenant colonel had—anachronistically—read recent game theory, he might have concluded that it was irrational to cooperate with Trifonov in the first place. Their interaction was de facto a finitely iterated prisoner’s dilemma. The usual argument against cooperating in such a game begins with the premise that one should treat the final play of a finite series of plays of the game as a one-shot game, in which one should defect. But if one should defect on the final play, then the penultimate play is de facto a final play in the sense that it can have no effect on anything thereafter, and so one should defect on the penultimate play as well. By tedious induction backward, one should defect already on the first play in the series.

If the backward induction argument is compelling, it is hard to see how rational individuals could ever enter into normal relationships of trust and exchange. All such relationships would have to be grounded in something extrarational, perhaps in normative commitments to be more decent than is rational. That there is apparently a great deal of trust in our lives then suggests that we are not rational. I think, on the contrary, that trust is eminently rational and that the backward induction argument is flawed. In brief, the flaw is this. Suppose I know that you are eminently rational and that you believe the backward induction argument. I also know that we could gain
substantially from entering a series of exchanges that must terminate, perhaps unhappily, at some distant future point. I can now wreck your backward induction by simply cooperating at our first encounter. You may now suppose I am irrational, or you may reconsider your induction. Either way, you may now decide it is in your interest to reciprocate my cooperation, so that we both benefit far beyond what we would have got from continuous mutual defection. (I think you must reconsider your induction because, if I can get you to cooperate by acting cooperatively, you could do as well with others. That is to say, you must agree that it would be rational for you to cooperate initially rather than to defect.)

Moreover, and more to the point here, if you think cooperation in finitely iterated prisoner’s dilemma interactions is irrational, you must wonder at your own tendency initially to trust those whom you do not yet know well. All our relationships with people are of ill-defined but necessarily finite duration. The backward induction argument recommends initial distrust and, further, continued distrust. This is surely a recommendation for slow death by abnegation. Whatever the apparent force of the backward induction argument for rarefied game theorists, it appears that actual people in going societies regularly take the risk of initially cooperating to upset that argument. Only for that reason do we have going societies.

Various social scientific accounts of trust take for granted that trust is rational in the sense of being based on empirically grounded expectations of another person’s (or an institution’s) behavior (Barber 1983; Luhmann 1980). In effect, we do a reasonable job of reading tendencies and trends, such as those exhibited by a stock market or a large crowd of people such as doctors or politicians. Those who see trust as normative or otherwise extrarational argue that it is more richly a two-part relation than this view implies. Trust involves intentional or motivational moves by the trusted as well as by the trustor. A rational analysis of trust of another intentional being, as opposed to trust of a force of nature (our trust in the sun’s rising tomorrow), must attend to the rationality of both intentional parties. In this sense, sociological expectation accounts are only half-rational and they therefore fail to address the apparent concern of many moral philosophers. They have a liability not unlike that of the similarly half-rational Cournot theory of market behavior. A Cournot actor assumes regularity of behavior on the part of others in the market, but fails to take account of second-order effects of others’ responses to his or her own actions.

A fully rational analysis of trust would not depend solely on the rational expectations of the trustor, but also on the commitments, not merely the regularity, of the trusted. How can one secure commitments from someone whose love or benevolence does not guarantee good will toward oneself? The most common way is to structure incentives to match the desired commitment. You can more confidently trust me if you know that my own interest will induce me to live up to your expectations. Your trust then encapsulates my interest; it is a two-part intentional relation. On this view, as Schelling notes, “Trust is often achieved simply by the continuity of the relation between parties and the recognition by each that what he might gain by cheating in a given instance is outweighed by the value of the tradition of trust that makes possible a long sequence of future agreement” (Schelling 1960, pp. 134–135).

Several types of behavior often explained as moral can be clearly understood as self-interested. Promise keeping, honesty, and fidelity to others often make sense without any presupposition of a distinctively moral commitment beyond interest. Consider promise keeping, which has been the subject of hundreds of articles and books in moral theory during this century. Hume says, without seeming to think the statement requires much defense, that the first obligation to keep a promise is interest (Hume 1739–40, p. 523). The claim is obviously true for promises between close associates who have an ongoing relationship that they want to maintain. If I promise to return your book, I’ll be encouraged to do so by frequent contact with you and frequent desire to make other exchanges with you. If I generally fail to keep such promises, I can probably expect not to enjoy as many exchanges and favors. Promising relationships typically are those in which exchanges are reciprocated over time. Because exchanges are resolutions of prisoner’s dilemma problems (Hardin 1982a), promising relationships have the incentive structure of iterated plays of the prisoner’s dilemma.

It is sometimes supposed, on the contrary, that promising is typically used to regulate relations with strangers. Unless my worldly experience is extraordinary, this view is prima facie false. Promises to genuine strangers are rare, not least perhaps because they would not be trusted. Schelling canvasses peculiar devices for securing compliance with promises in such difficult contexts as those between
strangers (Schelling 1989). Establishing trust in such contexts requires strong measures, such as subjecting oneself to risk of real, often unrelated, harm if one fails to comply with one’s promise. When we have to trust strangers in important matters, we commonly prefer to bind them through contracts under law.

The force that generally backs promises is the loss of credibility that follows from breaking one’s promises. Without credibility, one loses the possibility of making promises. This sounds suspiciously similar to Schelling’s right to be sued. Why should anyone want the power to make promises? All I really want in my own interest is the power to receive them. And there’s the rub, because promises are generally reciprocal. The real penalty here, as in Schelling’s case if there were no right to be sued, is that others will no longer rely on me, but that they won’t let me rely on them (D. Locke 1986, p. 574). As is trust, promising is typically a two-part intentional relation.

Trust and promises are morally kin, just as much of the vocabulary of the two is cognate in its etymology. As with promising, future expectations, generally based in ongoing experience, contribute much of the force that binds in a trusting relationship. Trifonov and the lieutenant colonel could trust one another while future expectations of their relationship were motivating. Reputational effects, which can often put dishonest merchants out of business, were of no use to the lieutenant colonel because he himself could not afford to go public.

As this example illustrates, trust does not merely apply to another person but to another person under certain conditions. The particularity of many claims of trust can be explained by the particularity of interests at stake. I once had an acquaintance of whom many people said, with genuine force, that he was a person you could trust. Alas, that depended on who “you” were. Many people did not trust him at all because they thought him deceitful and manipulative. The latter group included people whose interests often conflicted with his and whose future value to him he had seemingly written off. He could be richly and deeply trusted by those who shared enough of his interests, not at all by those who did not. He was almost mythical in his capacity to put people into two distinct classes. Most of us are not mythical, but we can still make distinctions. Without a legally binding contract I might readily trust you to return the $10 I lend you for lunch, but quail at the thought of trusting you with $10,000—let alone 4,500 rubles that are not mine anyway. In Anna Karenina, Count Vronsky’s code of social rules was probably well understood by all concerned, who therefore had differential grounds for trust and distrust. As Tolstoy puts it, “The code categorically determined that though the card-sharp must be paid, the tailor need not be; that one may not lie to a man, but might to a woman; that one must not deceive anyone, except a husband; that one must not forgive an insult but may insult others, and so on” (Tolstoy 1949/1875–77, book I, p. 347). Karenin therefore could trust Vronsky with a gambling debt—but not with his wife.

Many of the things that one could trust Vronsky to do are different from the kind of sequential exchanges that can be modeled by the iterated prisoner’s dilemma. Someone’s trust in him would be a two-part intentional relation, because it would depend on Vronsky’s intentional and arguably rational commitments. (They are simply rational if Vronsky adheres to the code just in case it fits his interests to do so. In choosing among debts to pay, he pays his fellow noble gambler, who could harm him socially or otherwise, rather than his tailor, who can only hassle him and refuse him further service.) Similarly, I can trust you to do what is in my interest in certain contexts because your doing so will fulfill some interest of your own. You need not necessarily give me something in return for something I give you. But the strategic relationship may be essentially the same as that for direct exchange. When I trust a political official who may be held at least somewhat accountable for failing to fulfill my interests, we are related in our intentions, though we may never meet.

Let us push the encapsulated interest view to the extreme. In Shizuko Go’s Requiem, a painfully beautiful novel about the destruction of a vast web of social relationships through wartime deaths in the last months of World War II in Japan, the heroine Setsuko recalls “the familiar precept of perfect hospitality: ‘We meet but once’” (Go 1985, p. 107). There is strategic subtlety in this bit of popular wisdom. If I know we meet but once, my hospitality is not my half of a two-part relationship, not an initial move in a potential trust or exchange relationship. It does not encapsulate your interest in reciprocity over the long run. It is purely a gift or an expression of my hospitable character.

In Setsuko’s case, the hospitable older woman whom she meets but once has a son whom Setsuko wishes to visit. Because the woman loves her son and wishes him well, interest should incline her to be nice to Setsuko on her singular visit. At the same time, independently of her interest, she may also be normatively motivated to kindness.
One could construct arguments for the rationality of developing strong normative commitments: for example, following a norm saves on decision costs in many contexts. But in many cases the behavior may simply be normative outright. Setsuko’s older woman has been taught to be kind in certain circumstances of hospitality, and she might behave that way more or less independently of broad incentives to vary her degree of kindness. She has simply made a virtue of hospitality. She might also have made a virtue of trusting people. In both cases, she would presumably conclude that particular people whom she has tried and found repeatedly wanting are not worthy of hospitality or trust. Still, her initial stance is one of virtue rather than of calculated self-interest.

There are other possibilities. You might extend hospitality or trust in order to demonstrate to me that you have faith in my morality or character, to give me an opportunity to live up to your hopes even though I may have no incentive to reciprocate your action. Or you may not wish to be the kind of person who acts toward another with distrust that is not based on solid evidence. Such motivations are apt to lead to disappointment in many contexts, but they might be statistically justified in certain milieus. In particular, they might be justified in contexts in which there are rich possibilities of further interactions. In such contexts, however, interest is likely to conspire with your hopes in getting me to reciprocate.

Trust as Ungrounded Faith or Belief

A common dictionary definition of trust reflects the sense that it is somehow more than mere expectations. As Webster’s New World Dictionary (1980, p. 1201, under “rely”) puts it, “to trust is to have complete faith or assurance that one will not be let down by another [to trust in God].” Abraham evidently had such complete faith in God as to be willing to sacrifice his beloved son merely on God’s order. But apart from such trust in God, which may amount to a blanket acceptance that whatever God causes to happen must be for the best, it is hard to imagine anyone reasonably assenting “complete” faith in anyone.

Consider the infant who is not yet able to trust or to distrust, who depends on the actions of parents and others, and who merely accepts those actions. From this early relation “there gradually evolve attitudes which may be called trustful” (Hertzberg 1988, p. 316). Or, if that early relation is very bad or capricious, the child may develop an utter incapacity to trust. Because there is no question of choice in the matter, it is not sensible to say that the infant trusts its parents. Indeed, it may not be sensible to say that Abraham trusts God, given his beliefs. He might have failed to follow God’s orders because of weakness of will or he might have revised his view of the goodness of God and the rights of following his orders. But if Abraham could not revise his beliefs, then there was no question of choice for him any more than if his son had been taken from him by disease.

Given the way in which trust seemingly develops from infancy, one might suppose that trust “is not based on grounds” (Hertzberg 1988, p. 318). The older woman whose kindness impressed Setsuko seems to have been normatively motivated. If so, then her kindness to Setsuko was not necessarily based on grounds. But then it would be odd to say of the particular people to whom she was gracious that she trusted them to reciprocate. Her graciousness was almost entirely an expression of herself, without objective correlates, not specifically directed at particular people. If trust is selectively directed at only certain other people, but not based on grounds, then it must be capricious—unrelated to its objects and not a consistent expression of character. Such trust seems neither sensible nor meritorious.

The philosopher Annette Baier writes, “What is the difference between trusting others and merely relying on them? It seems to be reliance on their good will toward one, as distinct from their dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will toward one, or on motives not directed on one at all” (Baier 1986, p. 234). Is there really a difference here? I rely on you, not just on anyone, because the experience that justifies reliance is my experience of you, not of everyone. Moreover, I can see you having good will toward me because it is generally in your interest to do so and against your interest to have ill will toward me. Perhaps you are an official in some organization or you have an ongoing relationship with me, either directly or indirectly through others. You benefit from that role or that relationship and you will want to maintain it through appropriate good will toward me. Your good will is probably quite restricted—there are many things you could not be expected to do on my behalf. My trust in you will probably also be restricted.

In some cases, you may have good reason to believe that someone loves you or has strong altruistic motivations toward you or people
like you. Your trust still turns on a rational reading of that person’s intentions. Trust does not depend on any particular reason for the trusted’s intentions, merely on credible reasons. In fact, many of our trusting relationships, especially those of early life, are grounded in the love or altruism of others. That may be one reason why we associate moral qualities with trust.

On Baier’s account, trust is an extension of the infant’s relation to its parent, especially its mother. Yet, she notes, a “constraint on an account of trust which postulates infant trust as its essential seed is that it not make essential to trusting the use of concepts or abilities which a child cannot be reasonably believed to possess” (Baier 1986, p. 244). If this is so, then the encapsulated interest view of trust is inherently wrong because assessments of trustworthiness could only be based on instinctive, behavioral learning. They could not require straightforward rational accounting such as we indulge regularly when, for example, we revise a prior supposition and decide someone is, after all, not trustworthy. On the infant view, trust is a primitive and somewhat ineffable condition in which we sometimes find ourselves. Such trust cannot very well apply to people whom we know almost entirely through intellectual apprehension; yet people say they trust public figures, merely on the basis of what they have read about them. There surely is some element of the primitive and ineffable in many of our commitments, perhaps especially in the forms they can take. In particular, our capacity for trust must build in part on evolved instincts. But our trust itself is not necessarily as primitive as the “innate readiness of infants to initially impute goodwill to the powerful persons on whom they depend” (Baier 1986, p. 242). That readiness may be a necessary or at least important foundation on which the capacity for trust may be built.

Consider an adult instance of Baier’s kind of trust. In Wagner’s opera Lohengrin, Lohengrin is an utterly incredible, godlike figure who demands of Elsa that she trust him without doubt or query or all is lost. Elsa is a true Wagnerian heroine, prepared to submit to and adore her hero as her lord. She wants not only to trust him, but to marry him, so beautiful a person does he seem to be. By refusing to tell her why she should trust him, Lohengrin puts her in the relation of an infant to her all-powerful parent with no choice but to accept or perish. Elsa’s fundamental problem is that she has no good way to explain Lohengrin’s existence and powers. He has come from nowhere, no one has ever heard of him. The nearest theory available to

Elsa for understanding him is sorcery. That theory would make Lohengrin evil, not good. Given her understanding of the world, it would be stupid of her to trust him merely on his demand. In the end, it is hard for us mortals to avoid thinking of him as inhuman and partly evil in his supreme goodness. The view of trust that Lohengrin imposes is repulsive.

An account of the life of the infant and its necessary dependency might seem cogent as an account of how we come psychologically to be able to trust or to know. But that is not an account of what trust is or of how it works. It is not only preadult, it is prehuman. It is plausibly the way an infant bird works, turning a wide-open mouth to the sky with an instinct behaviorally equivalent to trust that good things will fall into it. Such instinctual considerations are arguably a compelling part of an account of knowledge, including knowledge that backs up trust, knowledge of the reliability of any particular other and of others in general. That kind of knowledge is inherently inductive, and one might suppose it wise to be skeptical of inductive knowledge until the run is fairly long. But our normal proclivities may be to make an optimistic assessment of a short but so-far positive inductive run. If we live in a culture in which that optimism typically is justified by longer-run experience, we have rich relationships from trusting others. Still, the act of trusting, though it may more nearly befall us than be chosen by us, is one that depends on objective data and is subject to correction if the data recommend.

The most significant sense in which trust may go beyond justified expectations is that many of us—more, no doubt, in some societies than in others—face a new case with optimism, with tentative trust. But we are not wildly irrational in our optimism, and we will withdraw our tentative trust if it proves to have been unwarranted. This is a minimally rational constraint on trust: One will not continue to trust another who repeatedly fails the trust. Moreover, we may be more optimistic toward new cases in richly structured than in amorphic contexts. An account of the life of the infant child and its inherent dependency suggests the plausibility of evolutionary selection for openness that enables us later to be tentatively trusting.

Howard Margolis notes the social nature of most knowledge, a thesis that is well developed by Wittgenstein. “By and large the easiest and even the most reliable reason for believing X is to be aware that everyone else believes X” (Margolis 1987, p. 135). Trust in the knowledge of others may well have been a trait favored by natural
selection in earlier times. And it must be reinforced for us by the way in which we are educated to understand our world. As children we had little alternative to following advice that could not be grounded in anything other than authority, and we were often rewarded by success when we followed it. For these and other reasons, as Margolis (1987, p. 45) notes, cognition is “intrinsically ‘a-logical.’” We build our cognitions from patterns and from society, and in many cases we could no more justify them than we could prove basic laws of physics.

Trust as Inherently Moral

Trust is clearly a two-part relationship, although there is disagreement on what the parts are. On the encapsulated interest account, if I trust you, I have certain expectations about your behavior under various contingencies. Presumably those expectations are grounded in your behavior or apparent commitments. This notion can be augmented in at least two ways. First, my trust may be something about my character well beyond mere expectations of your contingent behavior. This appears to be Baier’s view (Baier 1986). Second, some discussions of trust treat it, explicitly or implicitly, as an essentially normative concept. If you do not live up to my trust, I may conclude that you have violated my expectations in a way that merits moral censure, as though you were obligated by my trust or by whatever gave rise to it.

In many of our relationships, trust is perhaps a bit like altruism. I trust you more than can be enforced by withdrawing from future interactions with you, a weak sanction. My trust is virtually a gift. Or, better, it is a gamble. You may live up to my trust, and we may then go on to have a strong and mutually rewarding relationship. But you may also turn out to be a Trifonov who will turn on me when the moment for gain is ripe. One might go further, as Jon Elster does, to say that “Altruism, trust and solidarity are genuine phenomena that cannot be dissolved into ultra-subtle forms of self-interest” (Elster 1979, p. 146). This clearly makes sense for altruism because I can genuinely have your interests at heart independently of any causal connection back to my own interests. Most of us are probably altruistic to some extent, even if not to a very great extent. But it is not clear what is analogous about trust in Elster’s view.12 Many writers and perhaps most people in ordinary life seem, with Elster, to have some vague sense of a distinctively moral character to trust. But, in contrast to the case of altruism, it does not make sense to cut trust free of mooring in expectations and, hence, at least potentially, in interests. We cannot cut it free because our expectations will be grounded in factual assessments of the motivations of anyone we might trust, among the most important of which must be interests.

Of course, ascriptions of trust may be normatively loaded in some ways, just as expectations and other interpersonal terms may be. For trust to have normative bite, however, it must entail some degree of obligation. That is, it must depend on something specifically relational as, for example, a contract does. The trusted must do something that morally motivates the claim of obligation. I trust you because we have some kind of relationship or because, at least, you typically have some kind of relationship toward those in a relevant class. For example, I trust you because we have been through a lot together, or because you are a police officer and I am a citizen. Our ongoing relationship or our role relationship may generate mutually reinforcing expectations that each of us sees as obligating to some extent and that each of us may have reason to think the other sees as obligating.

“When someone’s trust has been misplaced,” the philosopher Lars Hertzberg writes, “it is always, I want to say, a misunderstanding to regard that as a shortcoming on his part. The responsibility rests with the person who failed the trust.” He goes on to say that, “unlike reliance, the grammar of trust involves a perspective of justice: trust can only concern that which one person can rightfully demand of another” (Hertzberg 1988, p. 319; Hertzberg’s emphasis).13 On this view, I can only “trust” someone to do what that person already has a moral obligation to do. This is a definitional move that, if accepted, makes unnecessary Hertzberg’s several pages of argument while raising many ancillary questions. On whose moral theory do I ground your obligation, one might wonder, yours or my own? Hertzberg’s redefinition also makes “trust” a nearly otiose category.

Perhaps we could rescue the sense that trust is inherently moral by reserving the term “trustworthiness” for those whose reciprocity is motivated by character or morality rather than by interest. One might then suppose it plausible to speak of trust only in cases of trustworthiness and to speak of reliance when reciprocity is grounded in interest. Unfortunately, this restriction would imply that we generally cannot be sure whether we trust or merely rely because the data on another’s character, morals, and interests are likely to be confounded. Among
close associates, indeed, the bulk of the data are likely to be the same. But it would be absurd for the lieutenant colonel to say he only thought he trusted Trifonov, that he now realized he had not trusted him at all. He had trusted him, and therefore he had been swindled by him.

When our trust proves to be misplaced, we are likely to say, "But I trusted her." Indeed, we are likely to say this in many contexts in which we could not truthfully say, "But she gave me a commitment." Our trust depends on actions of the other but it is commonly not authorized by those actions, as it might be in the case of a contract or promise. In its common understanding, therefore, trust is not essentially normative, it does not imply an obligation on the part of the trusted, who may not even be aware of the trust placed in her. Hertzberg seemingly moralizes the notion of trust by simply defining it as coupled with assumed obligations. His concept of trust may still be a two-part relation, but it is a very one-sided relation with all of the burden on the trusted.

Unless one makes something like the definitional move that rules out the application of trust to any cases but those of moral obligation, it does not seem likely that trust can be moralized. It is an attitude that can be grounded in moral obligations, as public officials may be regarded as morally obligated to behave in certain ways toward their constituents. But it need not be. And it can be grounded in expectations about the moral commitments of others. But, again, it need not be.

In this respect trust is clearly different from promising and many other moralized notions. We may say that someone who makes a promise is obligated by it. This is merely a conventional, not a logical claim. There is a long tradition of claiming that to break a promise is to be logically self-contradictory. To make a promise is to proclaim an obligation; to break it is to prove that proclamation false. In more recent writings, the logical entailment of an obligation to keep a promise is tricked up out of the meaning of the ordinary words 'I promise,' which are taken to entail fulfillment. If we reject such an analysis, we may still suppose that competent people who make promises thereby assume some obligation, which, however, may fall short of actually requiring them to keep all their promises. But with trust we cannot even formulate an analogue of either the conventional or the logical account of an obligation. It is I who promise and who thereby assume an obligation. But if it is I who trust, it is the trusted who would have to be burdened with an obligation.

The Social Grounding of Trust

Those who use the term "trust" readily apply it to institutions and institutional actors, such as banks, nations, and political leaders. Many plausible psychological and normative accounts of individual behavior may be hard to generalize to apply to institutional behavior. A theory of trust that does not generalize to institutions is of limited interest in political theory and international relations. As observers of politics we often speak in analogies that may be fallacies of composition. For example, we may try to explain peaceful Anglo-American relations by saying England and the United States trust each other.

If our notion of trust comes from understandings of individual behavior and character, the term may be entirely out of place in application to a nation, group, or institution. There may be ways to interpret the notion to apply it to such actors, but it is not likely to be prima facie applicable without interpretation. It is now a commonplace understanding that interest does not readily generalize from individual to group or national levels. It should not surprise us to find that trust, which is commonly at issue just because interests are at stake, does not readily generalize either. Nevertheless, the encapsulated interest conception of trust can be generalized to fit institutions.

Advocates of moralized conceptions of trust at the individual-level argue that trust is inherently a two-part relation, that it is not merely rational expectations about the behavior of others. Sociological accounts of trust, such as that of Bernard Barber (1983), seem, however, to account for trust as simple expectation grounded in large-number regularities. Barber's interest in trust grows in large part out of his more general concern with the role of professionalism in our lives. We cannot know enough to judge the competence of the professionals who serve us; therefore, we must essentially trust them to some extent. Niklas Luhmann (1980) focuses on our need to trust (that is, to have stable expectations about) large institutional aspects of life, such as the stability of our currency or the reliability of our political leaders during crises that could lead to foolish war.

On these accounts, large-number stability in the behavior of relevant others or other types can be a reason for expecting more of the same behavior. Large-number regularities can play a stronger role: They can affect one's incentive to act as others do. My expectations of your behavior may turn in large part on my expectations of behavior of people like yourself in our society. For example, consider marital
fidelity. I might suppose that the best norm for marriage is to be utterly faithful until death do us part. If all followed that norm, I might firmly believe, ours would be a better world. Suppose the woman I wish to marry believes the same. We therefore publicly vow to be faithful forever. Should we believe each other? Perhaps. But only in the sense that we really are committed at this moment. Whether we will still be committed in a few years may depend very much on the laws and norms of our society.

In Verdi’s *La Traviata*, Germont wishes to discourage the former courtesan Violetta from continuing her life with his son, Alfredo. Violetta says that, alas, she can love no one but Alfredo. Changing his tack, Germont points out that “men are often fickle.” Violetta involuntarily says, “Oh God,” and Germont, given the opening, demolishes her hope. He sings:

One day, when time has dispelled  
the charms of love,  
tedium will set in quickly.  
What happens then? Think—  
The deepest feelings  
can bring you no balm,  
*since heaven has not blessed*  
this union. (Verdi 1846/1853, Act 2, Scene 1, p. 9)

The blessing of heaven was, of course, to be secured from earthly institutions that would then protect Violetta’s claims on Alfredo by, at the very least, making it illegal for him to marry anyone else. Like Schelling’s contractor, Violetta understood that restricting one’s freedom might be necessary to securing desirable ends. Of course, as with Schelling’s contractor, her greater concern was to restrict someone else’s freedom to secure desirable ends.

Recall also Vronsky’s code of ethics. Given the prevalence of this code among his class, Vronsky was virtually incapable of convincing anyone he would behave otherwise toward anyone’s beautiful wife than he behaved toward Karenin’s Anna. In this respect, trust, like expectations and knowledge, is in part inherently a social construction. Even if they were trustworthy, Vronsky and Alfredo would have no way to convince others.

If divorce is impossible and infidelity is severely sanctioned, we who wish to swear fidelity until death may find it quite easy to trust one another. If infidelity and opportunities for it are rampant, however, we may be unable to trust one another with anything short of an act of faith. In many contexts your capacity to make a commitment about how you will behave depends on the normal expectations of behavior. You cannot easily establish a commitment to a much more demanding standard of behavior. Genuinely normative commitments may depend on or interact with incentive structures.

Society offers two general categories of controls to individuals who can benefit from constraining themselves. There are the elaborate, large-scale controls of the law in a relatively extensive society. And there are the particular, small-scale controls of ongoing relationships of family, friends, and what we might crudely call geographical associates—those with whom we will almost inescapably be thrown into further dealings. Between these two categories there are institutionalized religious controls that blend elements of both. Where social controls fail we are left with our own personal devices of internal motivations.

One of the best and most useful of internal motivations is interest, which leads us to and through much of the best that life has to offer. To suppress interest, so that it does not lead us in ways we might consider wrong, is to put ourselves at war with ourselves. In waging that war we are supported by social controls based in our own long-term interest. A strong network of laws and conventions is needed to make any kind of behavior reliable if it is likely to conflict with powerful considerations of interest. Neither the law nor conventions worked to secure faithfulness from Vronsky and Alfredo. And few of their peers would expect faithfulness from them.

**Trust Institutions**

On the encapsulated interest account, trust and cooperation are related problems. They are not always the same problem. Cooperation may generally require conditions that make for trust, but not all trusting relationships are sensibly grounded in ongoing cooperation. Some trusting relationships depend on love or altruism from the trusted, and some involve only a loose concatenation of interests, nothing like direct exchange or pursuit of common goals. The latter may be especially typical of political life, where it may be vital to establish something analogous to trust for institutions, so that citizens may prosper and institutions and nations may cooperate.

Some, including the usually very sensible Hume (1978/1739–40, p. 537), have supposed that government requires public-spirited
people to make it work well. In practice most political institutions are staffed by individuals whose motives are heavily if not entirely self-interested. To gain our trust, they will have to work in our interest. Hence, the general problem is to make it the interest of various officials to work in our interest. We do this in part by making some officials directly answerable to citizens and in part by making other officials answerable to these. Both these controls are likely to be very loose, but the latter sounds especially weak. What we need to complete the picture is a theory of how the general interest can be served by a government of millions of bureaucrats who are fundamentally self-interested, who are motivated not by unusual public spirit, but only by income and career.

In crude outline, the most plausible theory is one that takes James Madison’s analysis down to the level of individual officials. In defense of the U.S. Constitution, he writes, “In framing a government which is to be administered by men over men, the great difficulty lies in this: you must first enable the government to control the governed; and in the next place oblige it to control itself.” He recommends, “Ambition must be made to counter ambition” (Madison 1961/1788, p. 322). How? If I violate the norms determined by our bureaucratic mission, you and others are likely to find it in your interest to oppose me (Hardin 1988a, pp. 526–527). Sometimes the enticements to malefeasance are so great that they infect almost everyone in a relevant agency, as we often hear of police units that succumb to bribery or even direct involvement in such profitable crimes as illegal drug trade. But commonly, even in such extreme cases, someone will have a strong career interest in bringing them to account. Strong moral commitment beyond interest may help and may be common, but it may also lead officials into taking the law into their own hands, and it cannot be reliable.

Often we expect institutions to be more stable than individuals. Many institutional promises and threats are more reliable than their individual equivalents. Consider some examples. The nuclear deterrent threat is credible because it does not depend on a particular individual’s commitment to act in the relevant moment. As has widely been discussed, the individual might choose, once deterrence had failed, to act as a humanitarian rather than as an avenger or an automaton programmed to retaliate. The reliability of retaliation approaches certainty if it is institutionalized in the form of many actors prepared to act in related but not entirely centrally controlled ways. Similarly, many of us might trust our fortunes to a bank more readily than to most individuals, perhaps including close friends and relatives with whom we expect our relationships to last our lifetimes.

Aristotle argued that the best of all governments would be a good monarchy and the worst would be a bad monarchy, or tyranny. The differences between a good and a bad democracy would be less great. To reduce variance, we might choose democracy as the preferred form of government in general. There may be little or no empirical study of Aristotle’s factual claims, but they sound sensible. Government by the many may induce a kind of regression toward the mean and hence much greater predictability.

Seldom in history has anyone gone so far toward establishing institutional trust as has Soviet President Mikhail Gorbachev, head of a system that has exhibited extraordinary variance. He has made some previously possible Soviet threats virtually impossible by putting institutional barriers in their way. For example, in inviting the reformation of the East European regimes and the dismantling of the iron curtain, he has greatly reduced the possibility of a sudden Soviet conventional attack on West Europe. By withdrawing troops and certain materiel he has made it virtually impossible to launch a secret attack without first visibly warning of attack during the necessary restoration of troops and equipment to the European theater. The obstacles he has created consist of institutional structures that can impede individual audacity.

Such institutional arrangements are appealing partly because they stabilize our expectations. Institutional behavior regresses toward the mean to average out the variance of individual behavior. Our expectations may not be grounded in any theory or explanation of why they are justified, but simply in experience. For example, political trust of many kinds may be easier in the Soviet Union now than it was a generation ago, although many of the older generation may still be reticent in trusting others with their opinions. And it must be harder in China now than it was shortly before the June 1989 massacre in Tiananmen Square.

One might suppose that trustworthy (that is, reliable) institutions are reliable because the right people are in the right places in them. But banks and many other institutions do not very rigorously select people for their roles, and it seems unlikely that reliability emerges from simple goodwill on the part of individuals in those roles. Most of us are somewhat like bank tellers: we are secured in our normal
honesty by institutional arrangements that make significant dishonesty risky, even difficult. Much of what looks like honesty is essentially self-interest at work.

Institutional arrangements may secure our expectations, and hence our trust, with the devices of self-interest, just as our own individual arrangements do. I become trustworthy by establishing a reputation and by setting myself up for real losses if I betray a trust. Those who fail to learn such lessons are seen as capricious and adolescent. Public officials and institutions also must live by the reputations they establish.

John Dunn has argued that the failure of political philosophers to consider trust is a source of weakness in their theories (Dunn 1988; also see Dunn 1984). He evidently understands trust to be a form of encapsulated interest—he often speaks of “rational trust.” He is concerned to restore Locke’s view that society turns power over to its governors, “whom the Society hath set over it self, with this express or tacit Trust, That it shall be employed for their good, and the preservation of their Property” (Locke 1988/1690, p. 381). The relationship of citizens to government is one of trust, not one of contract, and it is the possibility of this relationship and its working that are to be explained. No matter how one comes down on the textual warrant in Locke for this view, it offers an insightful and compelling reconstruction of the otherwise incoherent move to justify government from a supposed grounding in contract or consent.

To bring trust into political theory requires a micro-level account of how government works at the macro level. This will largely be an account of rational expectations of what government and its agents are likely to do. But the expectations will be rational not because they extrapolate from current and past actions, as might be adequate for a sociological account of credible expectations. Merely institutionalizing government and the implementation of policies should lead to greater stability of expectations, and hence to greater trust in this sociological sense. To reach Dunn’s concern, the expectations must also be rational in the sense of depending on the rational commitments of officials. Rationally grounded trust in officials therefore requires that the officials be responsive to popular needs and desires. To have incentive to be responsive, they must be somehow accountable, most plausibly, perhaps, through competitive elections.

Many aspects of individual trust of political institutions deserve much more extended discussion. Institutions, for example, play a role in underwriting even interpersonal trust. As Hume says of contracts, if they “had only their own proper obligation, without the separate sanction of government, they wou’d have but little efficacy in [all large and civiliz’d] societies. This separates the boundaries of our public and private duties, and shews that the latter are more dependant on the former, than the former on the latter” (Hume 1978/1739–40, p. 546). Hobbes may exaggerate the extent to which powerful institutional sanctions are required for grounding trust and promises, but he is not radically mistaken.

Concluding Remarks

Many social and moral norms are primarily manifestations of encapsulated interest, as trust is. Because this is so, we can count on others enough even to take an unduly optimistic view of them. We can afford to be trusting in general until our trust proves to be badly misplaced. If, in contrast, we had come to be distrusting in general, the result would surely be far worse than what we have. Presumably, it would look much more akin to much of international relations, as in the Cold War between East and West, in which distrust often seemed to be the baseline and trust to be, until recently, an unreachable goal.

Would it make sense for an individual to go through life without initial trust toward new acquaintances? One might be always distrusting until reputational or other evidence recommended trust in a particular case, and no doubt some people are initially distrustful to such a degree. For many people, such a stance would be too tedious and would cost more in lost opportunities than it would save in avoided harms. To act according to the backward induction argument would not generally be in their interest.

Perhaps our willingness to open by tentatively trusting others, even if only in small ways, underlies a commonplace claim that even the market and other more or less purely exchange relations depend on a general level of trust. Some economists consider this general level of trust a public good that is nevertheless voluntarily provided by individuals through their piecemeal actions (Arrow 1972; Hirsch 1978, pp. 78–79). There may be less extrarationality here than seemingly meets the eye. If our expectations are stabilized at a high enough level of cooperativeness, we may finally be able to treat much of the behavior we expect to encounter as a relatively benign force of nature, just as microeconomic theorists of the market essentially do.
Under typical circumstances in large markets, I can have very stable expectations of fairly good results from my entering ordinary exchanges. My trust in "the market" may be like my trust in the sun's rising tomorrow. I will correct specific details of my trust when any dealer out there violates it. But otherwise I will treat each dealer as benign, at least in the sense of not malignant. If I share Adam Smith's view that most dealers are likely to share my interests (because they must serve my interests to serve their own), I may even think of them as positively benign. The public good of generalized trust, then, may not require a moral foundation. It may be little more than an encapsulation of the self-interest of all or most of us.

Notes

I wish to thank Richard Zeckhauser for causing this paper to be written, for using its topic to cajole me into meeting deadline promises, and for lively discussions and commentaries on the topic and earlier drafts of the paper. I also wish to thank participants in his seminar on the papers of this book at Harvard and participants in the informal Tuesday evening seminar on contemporary moral and social theory at the University of Chicago for their cogent comments on an earlier draft. In particular, Avon Leong, Jay Patel, and Alan Wertheimer wrote valuable, extensive commentaries. Finally, I thank Richard Arneson, Howard Margolis, Jonathan Riley, and Duncan Snidal for their comments. All of these people raised far more questions than I have answered.

1. Trifonov's misappropriated rubles thereafter thread their complex way through Dostoyevsky's entire novel, wrecking lives while motivating the plot.

2. For more extensive discussion, see Hardin 1982b, ch. 9. The conclusion of the backward induction argument has become a virtual dogma despite the fact that many, perhaps most, discusssants think it perverse. Its appeal as a dogma may simply be that it is cute and perversely contrary to common sense.

3. For more extensive discussion of promising in its strategic variety, see Hardin 1988b, pp. 41–44, 59–65.

4. Baier says that exchange of promises typically requires "one to rely on strangers over a period of time" (Baier 1986, p. 246).

5. The social psychologist David Good apparently shares my experience. He says, "Rarely is it the case that exchanges requiring trust are ahistorical single cases" (Good 1988, p. 33).

6. As Baier herself recognizes in her "Promises, Promises, Promises" (Baier 1985).

7. This is a large part of Baier's concern in Baier 1986 (see p. 236 and passim thereafter).

8. That precept is striking in the most ordinary circumstances, but it seems almost dreadful in the context of Setsuko's recollection of it. It was, she notes, "literally true of everything that happened now."

9. As suggested by Alan Wertheimer. Or you may act according to Gregory Kavka's Copper Rule of doing unto others what they do unto you, even when it is not strictly in your interest (Kavka 1986, pp. 347–348). Hence you may be trustworthy toward me because I am trustworthy toward you.

10. Hertzberg (1988) wishes to show a difference in the "grammars" of reliance and trust. He draws on discussions of trust by Wittgenstein. Unfortunately, he is misled in some cases by the English translation he uses, because the words Wittgenstein uses in German are as nearly equivalent to reliance as to trust, in Hertzberg's senses. Wittgenstein's words translated as trust by Denis Paul and G. E. M. Anscombe include glaubig hinnehmen and sich auf etwas verlassen (Wittgenstein 1972, §§159 and 509).

11. Locke supposed that atheists could not consistently be trusted because they would not fear ultimate retribution from God. On his account, then, trust is essentially a matter of rational expectations grounded in the rationality of the trusted.

12. Elster's concern is with altruism, and he does not spell out the analogous nature of trust implied in his brief aside.

13. Talk of the grammar of trust is risky. We might also look to its etymology, where among other things we find that it has an explicable common root with "tryst."


15. The success of a deterrent strategy may not require anything near certainty of retaliation, because even a small risk is too much to invite.

16. Dunn's view constitutes a major revision of Lockean thought. (See further, Laslett 1988, pp. 114–117.)

References


