II | Convention Refined

1. Common Knowledge

Agreement, salience, or precedent, we have seen, can solve a coordination problem by producing a system of concordant first- and higher-order mutual expectations. We need only imagine cases to convince ourselves that higher-order expectations *would* be produced. But how? What premises have we to justify us in concluding that others have certain expectations, that others expect others to have certain expectations, and so on? And how is the process cut off—as it surely is—so that it produces only expectations of the first few orders?

Take a simple case of coordination by agreement. Suppose the following state of affairs—call it A—holds: you and I have met, we have been talking together, you must leave before our business is done; so you say you will return to the same place tomorrow. Imagine the case. Clearly, I will expect you to return. You will expect me to expect you to return. I will expect you to expect me to expect you to return. Perhaps there will be one or two orders more.

What is it about A that explains the generation of these higher-order expectations? I suggest the reason is that A meets these three conditions:

(1) You and I have reason to believe that A holds.
(2) A indicates to both of us that you and I have reason to believe that A holds.
(3) A indicates to both of us that you will return.

What is indicating? Let us say that A *indicates* to someone x that

____ if and only if, if x had reason to believe that A held, x would thereby have reason to believe that ____. What A indicates to x will depend, therefore, on x's inductive standards and background information.

The three main premises (1), (2), (3), together with suitable ancillary premises regarding our rationality, inductive standards, and background information, suffice to justify my higher-order expectations. Let us see how my reasoning would work.

Consider that if A indicates something to x, and if y shares x's inductive standards and background information, then A must indicate the same thing to y. Therefore, if A indicates to x that y has reason to believe that A holds, and if A indicates to x that ____, and if x has reason to believe that y shares x's inductive standards and background information, then A indicates to x that y has reason to believe that ____(this reason being y's reason to believe that A holds). Suppose you and I do have reason to believe we share the same inductive standards and background information, at least nearly enough so that A will indicate the same things to both of us. Then (2) applied to (3) implies:

(4) A indicates to both of us that each of us has reason to believe that you will return.

And (2) applied in turn to (4) implies:

(5) A indicates to both of us that each of us has reason to believe that the other has reason to believe that you will return.

And so on *ad infinitum*, since each new conclusion begins "A indicates to both of us that . . ." Note that this is a chain of implications, not of steps in anyone's actual reasoning. Therefore there is nothing improper about its infinite length. Figure 26 is a more detailed representation of these implications in my case; those in your case could be represented similarly.

Consider next that our definition of indication yields a principle of detatchment: if A indicates to x that ____ and x has reason to

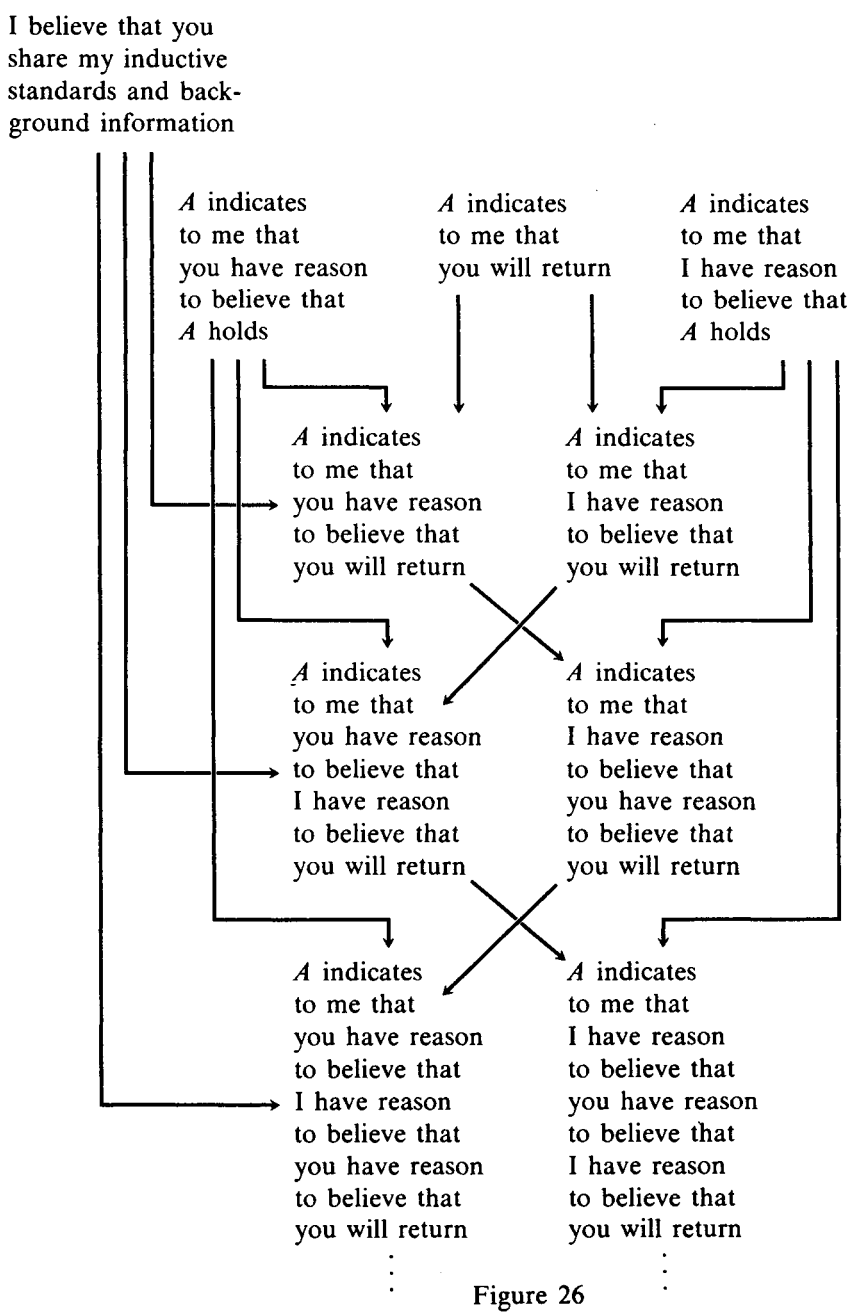

Figure 26

believe that A holds, then x has reason to believe that ____. Premise (1) applied in this way to (3) implies:

(3′) Each of us has reason to believe that you will return.

Premise (1) applied to (4) implies:

(4′) Each of us has reason to believe that the other has reason to believe that you will return.

Premise (1) applied to (5) implies:

(5′) Each of us has reason to believe that the other has reason to believe that the first has reason to believe that you will return.

And so on, for the whole infinite sequence we considered above. I am still not talking about anyone's actual reasoning or what anyone actually does believe. But the only actual reasoning needed now is reasoning to convert these iterations of "has reason to believe" to the corresponding iterations of "does believe." For that we need ancillary premises about rationality.

Anyone who has reason to believe something will come to believe it, provided he has a sufficient degree of rationality. So according to (3′), if we both have a sufficient degree of rationality, then it will come to be that

(3″) Each of us expects that you will return.

According to (4′), if each of us has reason to ascribe a sufficient degree of rationality to the other, then each has reason to expect that the other expects that you will return. If, in addition, we both have a sufficient degree of rationality, then it will come to be that

(4″) Each of us expects that the other expects that you will return.

According to (5′), if each of us has reason to expect that the other has reason to ascribe a sufficient degree of rationality to him, then

each has reason to expect that the other has reason to expect that he expects that you will return. If, in addition, each of us has reason to ascribe a sufficient degree of rationality to the other, then each has reason to expect that the other expects that he expects that you will return. And if, in addition, we both have a sufficient degree of rationality, then it will come to be that

(5″) Each of us expects that the other expects that he expects that you will return.

And so on. Each term of the sequence (3′), (4′), (5′) . . . , together with sufficient rationality, reason to ascribe sufficient rationality, etc., guarantees formation of the corresponding first- or higher-order expectation. But the degrees of rationality we are required to have, to have reason to ascribe, etc., obviously increase quickly. That is why expectations of only the first few orders are actually formed. The generating process stops when the ancillary premises give out.

This completes our example of a state of affairs which produces higher-order expectations. I take this example to be typical; all the higher-order expectations involved in sustaining conventions, and more or less all we ever have, seem to be produced in this way.

Let us say that it is *common knowledge* in a population P that ____ if and only if some state of affairs A holds such that:

(1) Everyone in P has reason to believe that A holds.
(2) A indicates to everyone in P that everyone in P has reason to believe that A holds.
(3) A indicates to everyone in P that ____.

We can call any such state of affairs A a *basis* for common knowledge in P that ____. A provides the members of P with part of what they need to form expectations of arbitrarily high order, regarding sequences of members of P, that ____. The part it gives them is the part peculiar to the content ____. The rest of what they need is what they need to form *any* higher-order expectations in the way we are considering: mutual ascription of some common inductive standards

and background information, rationality, mutual ascription of rationality, and so on.

Let us return to our example and consider the state of affairs A more completely. Suppose that as part of A we manifest our conditional preferences for returning to the meeting place. Then A may also indicate to us that we both have such preferences. If so, A can serve as a basis not only for common knowledge that you will return, but also as a basis for common knowledge that each of us prefers to return if the other does. Suppose also that as part of A we somehow manifest a modicum of rationality. Then A may indicate to us, and be a basis for common knowledge of, our possession of this modicum of rationality. By now A—our incident of agreeing to return—is generating all the higher-order expectations that contribute to our success in solving our coordination problem by means of replication.

A basis for common knowledge generates higher-order expectations with the aid of pre-existing higher-order expectations of rationality. Can these themselves be generated by some basis for common knowledge? Yes, because all the higher-order expectations of rationality needed to generate an nth-order expectation are themselves of less than nth-order. What cuts off the generation of higher-order expectations is the limited amount of rationality indicated by any basis—not any difficulty in generating higher-order expectations of as much rationality as *is* indicated by a basis.

Agreement to do one's part of a coordination equilibrium is a basis for common knowledge that everyone will do his part. Salience is another basis for common knowledge that everyone will do his part of a coordination equilibrium; but it is a weaker basis, in general, and generates weaker higher-order expectations, since the salience of an equilibrium is not a very strong indication that agents will tend to choose it. Precedents also are a basis for common knowledge that everyone will do his part of a coordination equilibrium; and, in particular, past conformity to a convention is a basis for common knowledge of a tendency to go on conforming. Consider a conventional regularity R in a population P. Everyone in P has reason to

believe that members of P have conformed to R in the past. The fact that members of P have conformed to R in the past indicates to everyone in P that everyone in P has reason to believe that members of P have conformed to R in the past. And the fact that members of P have conformed to R in the past indicates to everyone in P that they will tend to do so in the future as well.

For example, drivers in the United States have hitherto driven on the right. All of us have reason to believe that this is so. And the fact that this is so indicates to all of us that all of us have reason to believe that drivers in the United States have hitherto driven on the right and also that drivers in the United States will tend to drive on the right henceforth.

Our defining conditions for the existence of a convention consist of a regularity in behavior, a system of mutual expectations, and a system of preferences. I propose to amend the definition: not only must these conditions be satisfied, but also it must be common knowledge in the population that they are. Our amended definition is:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in any instance of S among members of P,
>
> (1) everyone conforms to R;
> (2) everyone expects everyone else to conform to R;
> (3) everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a coordination equilibrium in S.

Thus there is to be some state of affairs A (such that A holds, everyone in P has reason to believe that A holds, and A indicates to everyone in P that everyone in P has reason to believe that A holds) which indicates to everyone in P that members of P conform to R, that they expect each other to conform to R, and that they have prefer-

ences which make uniform conformity to R a coordination equilibrium.

One reason to amend the definition of convention is simply that we want to write into the definition all of the important features common to our examples, and common knowledge of the relevant facts seems to be one such feature. There is another reason: the amendment helps to deal with certain odd cases, regularities which seem intuitively unlike clear cases of convention but which would have qualified as conventions under the unamended definition.

Suppose everyone drives on the right because he expects everyone else to drive on the right and he wants to prevent collisions. But suppose no one gives anyone else credit for intelligence equal to his own. Everyone holds this false belief (call it f): "Except for myself, everyone drives on the right by habit, for no reason, and would go on driving on the right no matter what he expected others to do." This is a case of convention under the unamended definition, despite the false beliefs; but I think it ought to be excluded. It cannot be a case of convention under the amended definition (unless we are extremely irrational); for if it is, there is some state of affairs which we have reason to believe holds and which indicates to us that f is false. This case is only the first of a sequence. Suppose next that no one really has the false belief f, but everyone falsely ascribes it to everyone else. This too cannot be a case of convention under the amended definition (unless we are extremely irrational); for if it is, there is some state of affairs which we have reason to believe holds and which indicates to us that everyone has reason to disbelieve f. And so on. The cases become more and more unlikely, but no less deserving of exclusion; the amended definition continues to exclude them (given a sufficiently strong assumption of rationality—stronger and stronger assumptions of rationality are needed as we go on).

By now one might guess that common knowledge is the only possible source of higher-order expectations. But it is not; there is a general method for producing expectations of arbitrarily high order in isolation. For instance, I can acquire an isolated fourth-order

expectation as follows. Suppose I am a resident of Ableton and I believe everything printed in the *Ableton Argus*. Today's *Argus* prints this story:

> The *Bakerville Bugle* is totally unreliable; what it prints is as likely to be false as to be true. Yet the residents of Bakerville believe everything in it. Today's *Bugle* printed this story:
>
> > The *Charlie City Crier* is totally unreliable; what it prints is as likely to be false as to be true. Yet the residents of Charlie City believe everything in it. Today's *Crier* printed this story:
> >
> > > The *Dogpatch Daily* is totally unreliable; what it prints is as likely to be false as to be true. Yet the residents of Dogpatch believe everything in it. Today's *Daily* printed this story:
> > >
> > > Tomorrow it will rain cats and dogs.

I should not expect it to rain cats and dogs. I should not expect the residents of Dogpatch to expect it to rain cats and dogs. I should not expect the residents of Charlie City to expect the residents of Dogpatch to expect it to rain cats and dogs. But I should expect the residents of Bakerville to expect the residents of Charlie City to expect the residents of Dogpatch to expect it to rain cats and dogs. In other words, I should have a fourth-order expectation that it will rain cats and dogs, without any corresponding lower-order expectations that it will. Obviously, the method would have worked for an arbitrarily long sequence of newspapers; the sequence could have repeated, provided no two adjacent terms were the same. I do not claim that this method of generating isolated higher-order expectations is of much practical importance; it merely establishes the possibility.

2. Knowledge of Conventions

Suppose it is common knowledge in a population P that some state of affairs B holds. Then everyone in P has reason to expect it to be

common knowledge in P that B holds. For by definition of common knowledge, there is some state of affairs A such that:

(1) Everyone in P has reason to believe that A holds;
(2) A indicates in P that everyone in P has reason to believe that A holds;
(3) A indicates in P that B holds.

From (1) and (2) we may infer:

(4) Everyone in P has reason to believe that everyone in P has reason to believe that A holds.

From (2) by itself we may infer:

(5) Everyone in P has reason to believe that A indicates in P that everyone in P has reason to believe that A holds.

Likewise from (3) we may infer:

(6) Everyone in P has reason to believe that A indicates in P that B holds.

And from (4), (5), and (6) we may infer that everyone in P has reason to believe that there is a state of affairs A which satisfies conditions (1), (2), and (3).

So if a convention, in particular, holds as an item of common knowledge, then to belong to the population in which that convention holds—to be party to it—is to know, in some sense, that it holds. If a regularity R is a convention in population P, then it must be true, and common knowledge in P, that R satisfies the defining conditions for a convention. If it is common knowledge that R satisfies them, then everyone in P has reason to believe that it is true, and common knowledge in P, that R satisfies them; which is to say that everyone in P must have reason to believe that R is a convention.

This is not to say that a party to the convention has any special, infallible way of acquiring his knowledge. But he must *have acquired* it somehow, in an ordinary way, in order to be one of those among

whom the convention holds. Discovery of the convention is the principal part of one's initiation into it.

Consider the conventions of language, whatever they may be. Anyone who is a member of a population *P*, and party to its conventions of language, must know what those conventions are. If any regularity *R* is in fact a convention of language in *P*, any normal[1] member of *P* must have reason to believe that *R* satisfies the defining conditions for a convention.

Here is a vindication of sorts for Stanley Cavell's doctrine that a native speaker has no need of evidence to justify that he says about what he would say. Take a philosopher who claims we would not call an action voluntary if it were not abnormal. He need not cite occasions on which people have failed to call normal actions voluntary, for he is a native speaker of the language he is telling us about. Why is he excused? Not because he, as a native speaker, has some peculiar and infallible way of acquiring his knowledge of his language. And not because his knowledge of what we would say is not real knowledge, as Cavell seems to think when he says, "the native speaker can rely on his own nose; if not, there would be nothing to count." For the man who says what we would say is *not* just speaking for himself.[2]

Rather, it is because the knowledge Cavell has in mind is the speaker's knowledge of conventions to which he himself is a party. When Cavell speaks of our knowledge of "what we would say," I take it he means our knowledge of what we *could* say—could say without violating our conventions of language. He does not mean our knowledge of what we would say in order to provide our audience with the information they want; of what we would say in order not to be rude or boring; of what we would say in order not to divulge trade secrets; of what we would say in order not to twist our tongues.

[1]Not counting children and the feeble-minded, who may conform to *R* without expecting conformity and without preferring to conform conditionally upon the conformity of others.

[2]"Must We Mean What We Say?" *Ordinary Language: Essays in Philosophical Method,* ed. Vere Chappell (Englewood Cliffs, New Jersey: Prentice-Hall, 1964), pp. 75–112.

Once we have acknowledged that someone is a native speaker of our language, we have already granted that he is party to our conventions. Therefore he knows what those conventions prescribe; he knows "what we would say" in the sense in question. If we turn around and ask him to produce evidence for what he says about what we would say, we challenge his status as a native speaker and as a party to the conventions. We do not challenge some further status he might claim as an authority on the conventions *as well as* a party to them. He has evidence—perfectly ordinary evidence. But if we ask him to show it, we question his membership in the linguistic community to which he purports to belong. It makes no sense *both* to demand evidence for what he says about conventions *and* to take for granted that he is party to those conventions.

This vindication of Cavell's doctrine is a poor sort of vindication, however, because our knowledge of our conventions—that minimum of knowledge everyone has in virtue of his own participation—may be quite a poor sort of knowledge:

(1) It may be merely potential knowledge. We must have evidence from which we could reach the conclusion that any of our conventions meets the defining conditions for a convention, but we may not have done the reasoning to reach the conclusion. If asked whether something is a convention, we might give a snap judgment instead of evaluating our evidence; so we might get the wrong answer.

(2) It may be irremediably nonverbal knowledge. We recall the rowers in Hume's boat, example (3) of Chapter I.5. If I am one of the rowers who row in a certain rhythm by a tacit and temporary convention, I have evidence that we have a convention to row in that rhythm. Our success in rowing in that rhythm for the last few strokes is evidence by which I arrive at my expectation that you will continue to row thus; that you prefer to row thus if I do; and that you expect me to go on rowing thus. And it is evidence that you observe this same evidence. I can use such evidence, I can expect you to use it, and so on; but I cannot describe it. I cannot say how we are rowing—say, one stroke every 2.3 seconds—but I can keep on rowing that way; I can tell whether you keep on rowing that way; later, I could

probably demonstrate to somebody what rhythm it was; I would be surprised if you began to row differently; and so on. Now there *is* a description that can identify the way we are rowing. We take 1.4 ± .05 seconds for the stroke and .9 ± .1 for the return, exerting a peak force of 70 ± 10 pounds near the beginning of each stroke, moving the oars from 32° ± 6° forward to 29° ± 4° back, and so on, in as much detail as you please. But, as we row, we have no use for this sort of description. We can neither give it nor tell whether it is true if somehow it is given. We would need instruments, and even if we had them we could not go on rowing as we were while we took the measurements.

Like it or not, we have plenty of knowledge we cannot put into words. And plenty of our knowledge, in words or not, is based on evidence we cannot hope to report. Our beliefs are formed under the influence of impressions left by a body of past experience, but it is only occasionally that these impressions allow us to report the experience that created them. You probably believe that Kamchatka exists. Your belief is justified, for it is based on evidence: mostly your exposure to various books and to incidents that confirm the reliability of such books. Try, then, to make a convincing case for the existence of Kamchatka by reporting parts of your experience. There is no reason why our knowledge of our conventions should be especially privileged. Like any other knowledge we have, it can be tacit, or based on tacitly known evidence, or both.

(3) It may be knowledge confined to particular instances, taken one at a time. A regularity is conventional in virtue of certain general expectations and preferences regarding conformity to it. But these will not have to be general *in sensu composito;* generality *in sensu diviso* will suffice.

The distinction, Abelard's, is this. If I expect every driver to keep right, *in sensu composito*, then I have one expectation with general content: I expect *that* every driver will keep right. It does not follow that if Jones is a driver, I expect that *he* will keep right, for I might not realize he is a driver. Indeed, I might even realize that Jones is a driver and still not expect that *he* will keep right, for I might

fail to draw the proper conclusion from my general expectation. If, on the other hand, I expect every driver to keep right, *in sensu diviso*, then I have many expectations, each with *non*general content. I expect *of* Jones, a driver, that *he* will keep right. Of Morgan, too. And so on, for all the drivers there are. I need not know that Jones, Morgan, and the rest are all the drivers there are; I might falsely believe there are other drivers who do not keep right. Or I might altogether lack the general concept of a driver. Generality *in sensu composito* and generality *in sensu diviso* are compatible and often coexist; but it is possible to have either one without the other.

Generality *in sensu diviso* is problematic because expectation and the like apply fundamentally to states of affairs. If I expect that each driver will keep right, I do expect a state of affairs: each driver will keep right. But if I expect, *of* each driver, that *he* will keep right, what states of affairs do I expect? "*He* will keep right" does not specify *any* state of affairs until the pronoun has been replaced by some sort of description—verbal, pictorial, or otherwise—of the person in question. Suppose the description, "the driver of the puce Cadillac ahead of me," fits x. Then I can expect *of* x that *he* will keep right by having an expectation which attaches to x through that description of him: I expect that the driver of the puce Cadillac ahead of me will keep right. In that case, there is a state of affairs I expect. But not just any description of x will do. Suppose, unknown to me, x happens to be the chief of police and also the town drunk. I do not expect *of* x that *he* will keep right just because I expect that the chief of police will keep right. I do not fail to expect *of* x that *he* will keep right just because I do not expect that the town drunk will keep right. My expectation needs to be attached to x by a description of some special sort; and it is hard to say which descriptions will do, and why.[3]

Consider the general case: I expect every member of P involved

[3]See Quine, "Quantifiers and Propositional Attitudes"; Richard Montague and Donald Kalish, "'That'," *Philosophical Studies,* 10 (1959), pp. 54–61; David Lewis, "Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy,* 65 (1968), pp. 113–126; David Kaplan, "Quantifying In," *Synthese,* in press.

with me in an instance of S to conform to R. We have two universal quantifications: one over instances of S, another over members of P involved in any one instance. It is possible, of course, for my expectation to be general *in sensu diviso* over instances of S, but general *in sensu composito* over agents in any one instance. That is, it might be that I expect of any instance of S in which I am involved that everyone in it will conform to R. (Of course it is not possible for my expectation to be general *in sensu composito* over instances and *in sensu diviso* over agents.)

The same distinction between kinds of generality applies to other attitudes. Take our conditional preferences for conformity to convention—say, my preference for conforming to R in instances of S among members of P. If my preference is general *in sensu composito* over instances of S, then I prefer the state of affairs in which I conform to R whenever I am involved in an instance of S (among members of P who conform to R) to the state of affairs in which I sometimes fail to conform to R when I am involved in an instance of S (among members of P who conform to R). But if my preference is general *in sensu diviso*, then for any instance of S (among members of P) in which I am involved, I prefer the state of affairs in which I and the others conform to R in that instance to the state of affairs in which the others conform to R in that instance but I do not. Again the two kinds of generality can and often do coexist, but they are independent.

Which kind of generality over instances of S is wanted in the definition of convention? I should say: whichever kind it is that ensures the agent's ability to apply his general attitudes to the instance at hand. And that is a limited generality *in sensu diviso*. Whenever the agent finds himself in an instance of S among members of P, he must expect the others to conform to R *in that instance*, prefer to conform to R if they do *in that instance*, and so on, in order that he may have reason to conform to R himself. Attitudes general *in sensu composito* would be a likely and welcome addition, and could serve as a source of attitudes general *in sensu diviso*. But they would

not be enough by themselves; the agent would have to be able to recognize instances of S and derive the proper particular attitudes. If he did, his attitudes—or at least his propensity to acquire attitudes—would be general *in sensu diviso*.

We can imagine how a convention R regarding action in S might hold in a population P of creatures incapable of having any attitudes general *in sensu composito*. They learn from experience not by coming to believe generalizations, but by acquiring propensities to come up with the right particular beliefs regarding any new case that is presented in sufficient detail. They are exposed to precedents: what we would call (but they could not) instances of S, in which outcomes satisfactory to all concerned were reached by what we would call (but they could not) conformity to R. Thereafter, whenever one of them is presented with a new instance of S, even one not quite like any precedent, he has all the proper attitudes regarding that instance. He expects each other agent involved to do something we would call (but he could not) conformity to R. Considering any two outcomes, he has a preference; and his system of preferences between outcomes is such that there is a coordination equilibrium in which all concerned conform to R. Finally, it is common knowledge among members of P that these attitudes are present in each who is involved in *this particular* instance of S. There is a state of affairs A such that, for this or any other instance of S, for each one involved therein, A indicates that he has the appropriate attitudes in that instance.

Such a creature has a convention and knows it to this extent: given any instance of S, he knows how he and each of his fellows would act therein (namely, in some way that we would call conformity to R). And he knows that they do so by convention; that is, given any of the defining conditions of convention as applied to a given agent in the given situation, he knows the condition is satisfied. But he cannot think of more than one instance—the given one—at a time. He has no general concept of an instance of S, of a member of P, or of an action in conformity to R.

Suppose we who *do* generalize want to exploit this creature's

knowledge of his convention, in order to give a general description of that convention. We will have to proceed by trial and error, thinking up hypotheses and trying them out on him in one (well-chosen) instance after another until we think we can predict his response to any future instance.

Even we who could know our own conventions generally *in sensu composito* might happen to know them only generally *in sensu diviso.* If we wanted to know them generally *in sensu composito* as well, we would have to resort to the same sort of trial and error, with ourselves as subjects. Our data about instances of our own conventions would be reliable. But our general hypotheses to systematize those data would be ordinary tentative hypotheses with no privileged status.

3. Alternatives to Conventions

One of my defining conditions for the conventionality of a regularity R regarding choice of action by agents in a situation S has been:

> In any instance of S among members of P, everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a coordination equilibrium in S.

In the discussion of example (7) in Chapter I.5, we found this condition unsatisfactory whenever we had coordination between actions in nearby instances of S within some continuous activity, not just coordination between actions in any one instance of S. And we saw that the remedy was not to take longer stretches of activity as our coordination problems, for longer stretches are not coordination problems. Here I shall state new conditions that differ from the old one only by not requiring our activity to be chopped up into self-contained coordination problems. Our new conditions will not imply that S is a self-contained problem of interdependent decision, in

which each agent involved makes one choice of action and the outcome for each depends on the actions of all; but it will imply that *if S* is that, *then S* is a coordination problem and uniform conformity to *R* is a coordination equilibrium in *S*. The special case of a sequence of coordination problems will be covered as before; but we shall find that we have taken care of the other cases at the same time.

First, we require that each agent involved in an instance of *S* prefers to conform to *R* conditionally upon conformity by the others involved with him in *S*. He prefers uniform conformity to *R* to any combination of actions in which the rest conform and he does not. If *S* is a self-contained problem of interdependent decision, this first requirement makes uniform conformity to *R* an equilibrium. Otherwise, uniform conformity is not an equilibrium but something closely resembling one.

Second, we require that all agents involved have approximately the same preferences regarding combinations of their actions, so that *S* is a situation in which coincidence of interests predominates. In particular, we require that all share the conditional preference of each for his conformity to *R*. That is, just as I prefer to conform if you and the others do, you also prefer me to conform if you and the others do. Taking this and the first condition together: each prefers that everyone conform to *R*, on condition that at least all but one conform to *R*, whether that one is himself or someone else. If *S* is a self-contained problem of interdependent decision, this second requirement makes uniform conformity to *R* a proper coordination equilibrium.

Finally, we require that there is a second possible regularity *R′* (regarding choice of action by agents in *S*) which meets the same conditions we are imposing on *R*. We call *R′* an *alternative* to *R*. It is enough to require *R′* to meet the first and second conditions imposed on *R*. The third is automatic: if *R* has *R′* as an alternative, then *R′* has *R* itself as an alternative. If *S* is a self-contained problem of interdependent decision, this last requirement makes uniform conformity to *R′* a second proper coordination equilibrium. Thereby

it ensures that S meets the last condition defining a coordination problem: possession of two or more proper coordination equilibria.

Recall the discussion in Chapter I.2 of the triviality of any situation with a unique coordination equilibrium and predominantly coincident interests. We are now in a better position to describe this triviality: common knowledge of rationality is all it takes for an agent to have reason to do his part of the one coordination equilibrium. He has no need to appeal to precedents or any other source of further mutual expectations.

So far we have protected convention against this triviality by requiring S to be a coordination problem and hence, by definition, to have more than one proper coordination equilibrium. Now that we no longer require S to be a coordination problem, our requirement for an alternative continues the same policy. In fact, whenever S is a self-contained problem of interdependent decision, we have made no change at all.

Our new condition does serve to make evident one property of conventions that was not emphasized before: there is no such thing as the only possible convention. If R is our actual convention, R must have the alternative R', and R' must be such that it could have been our convention instead of R, if only people had started off conforming to R' and expecting each other to. This is why it is redundant to speak of an arbitrary convention. Any convention is arbitrary because there is an alternative regularity that could have been our convention instead. A convention that is *not* arbitrary, so to speak, is a regularity whereby we achieve unique coordination equilibria. Because it is not arbitrary, it does not have to be conventional either. We would conform to it simply because that is the best thing to do. No matter what we had been doing in the past, a failure to conform to the "nonarbitrary convention" could only be a strategic error (or compensation for someone else's anticipated strategic error, or compensation for someone else's anticipated compensation, etc.).

When we try to state the requirement for an alternative more carefully, a question arises. R and R' are supposed to be different,

which is to say that action in conformity to R (by an agent in S) is not also in conformity to R', and vice versa. But different always, or different sometimes? After all, instances of S do not have to be exactly alike. They merely have to be analogous, to fall under some common description that is natural enough to allow common knowledge of a propensity to extrapolate from some instances of S to others. So action in conformity to R might also be in conformity to R' for some agents in some instances of S, though not for all.

It is not good enough to require an alternative R' differing from R merely to the extent of being incompatible with R for some, or even for all, agents in some possible instances. Suppose S occurs in a frequent version, shown in Figure 27, and in a rare version, shown in Figure 28. (We neglect any further differences between instances

	C1	C2
R1	1 1	0 .5
R2	.5 0	.2 .2

Figure 27

	C3	C4
R3	1 1	0 .2
R4	.2 0	.5 .5

Figure 28

of a version.) S is trivial in one version but not in the other. Let R be the regularity of doing *R1* or *C1* (in the frequent version) or *R3* or *C3* (in the rare one). I take it that R ought not to qualify as a convention, since it is trivial in most instances of S. But it would qualify as a convention if we counted R' as its alternative, where R' is the regularity of doing *R1* or *C1* or *R4* or *C4*. R' is an eligible regularity that is incompatible with R in some instances of S.

It would be better to require an alternative R' that is uniformly incompatible with R, incompatible for every agent in every instance of S. Now R in the example above is disqualified. Its only uniformly incompatible alternative would be R'', the regularity of doing *R2* or

C2 or *R4* or *C4*. But R'' is not an alternative to R, since R'' usually fails to meet the requirement of conditional preference for conformity. In instances of the frequent version of S, no one wants to conform to R'' even if his partner does.

If every instance of S is a coordination problem, and if uniform conformity to R is always a proper coordination equilibrium, then we can find another proper coordination equilibrium in every instance of S. Hence the regularity R' whereby one does his part of a selected second proper coordination equilibrium in every instance of S is an alternative to R, and R and R' are uniformly incompatible.

If we prefer, however, we do not have to require a uniformly incompatible alternative to R. As a (seemingly) weaker version, we could just require that for every instance of S, there is a suitable regularity R' which is incompatible with R (for everyone involved) in *that* instance. Partially incompatible alternatives to R are good enough if there are enough of them. The two versions are not really different. The strong version implies the weak version directly; and the weak version implies the strong version indirectly, since we can always get a uniformly incompatible alternative by patching together pieces of partially incompatible ones.

Therefore we might replace our original condition by two new ones. This one:

> In any instance of S among members of P, everyone has approximately the same preferences regarding all possible combinations of actions.

together with either this one (strong version):

> There is some possible regularity R' in the behavior of members of P in S, such that no one in any instance of S among members of P could conform both to R' and to R, and such that in any instance of S among members of P, everyone would prefer that everyone conform to R', on condition that at least all but one conform to R'.

or this one (weak version):

> In any instance of S among members of P, there is some possible regularity R' in the behavior of members of P in S, such that no one in that instance of S could conform both to R' and to R, and such that everyone would prefer that everyone conform to R', on condition that at least all but one conform to R'.

I see nothing to choose between the two versions, and I choose the strong version for no good reason.

When S is a self-contained problem of interdependent decision, our new conditions agree with the original condition. But they do not require S to be self-contained. If not—as in my example of price setting, with S taken as a stretch of business activity long enough to include several pricing decisions—the new conditions are a natural extension of the original condition.

Let R be a convention regarding behavior in a coordination problem S; and let R' be another possible regularity, partially or uniformly incompatible with R, which would solve S. But suppose the coordination equilibrium we would reach by conforming to R' is much worse than the one we reach by conforming to R—so much worse, in fact, that it is only slightly preferred to some of the outcomes that are not coordination equilibria. Then do we really want to call R' a possible alternative convention? And do we want to say that R' contributes to the arbitrariness and conventionality of R? Perhaps not. Fortunately, our definition as it stands is likely to disqualify this R' as an alternative to R. It may be true that:

> In any instance of S among members of P, everyone would prefer that everyone conform to R', on condition that at least all but one conform to R'.

But it may not be true as an item of common knowledge. Our weak conditional preferences for conformity to R' may well fail to be indicated by any state of affairs A which we all believe to hold and which indicates to us that it holds. But of course we still require the satisfaction of the conditions to be common knowledge in P. And rightly: if our conditional preference for conformity to R' existed,

but not as an item of common knowledge, R' could not have sustained itself in the way a convention does, so it is not true that R' could have been our convention instead of R.

Consider a convention establishing some meeting place. Its alternatives would be the possible regularities whereby we would meet at other places. Some places are better than others. Some are so bad we would forgo meeting rather than go there. Considering worse and worse places, we come to the point where conditional preference for conformity fails: some of us would not want to go to the place even if the others were there. Common knowledge of conditional preference fails sooner: there are places good enough that each would want to go there if the others were there, but not good enough that we could count on each other to want to go there if the others did, count on each other to count on each other to want to go there if the others did, and so on. These places do not provide alternatives to our convention, so they do not contribute to the conventionality of our meeting place.

Or consider the conventions of our language. Their alternatives are the conventions of other possible languages. But how about a hypothetical language—or shall we call it a cipher?—so clumsy that even after any amount of practice we would still take minutes of paper-and-pencil calculation to construct or construe its easiest sentences? All of us *might* find even that language better than none, worth learning to use among others who used it. But if it were not common knowledge that we would, this language would not be among the alternatives that make our actual language conventional.

If the only alternatives to R were of this deficient sort, R would not be a convention. Neither would it have been a convention under the original condition requiring that S be a coordination problem, since that would not have been common knowledge either. Nor should it be called a convention. If the alternatives to R are such an inconspicuous feature of the situation, R seems almost as trivial as if they were not there at all.

Can R' be an alternative to R if the idea of acting in conformity

to R' has never occurred to anybody in P? The principle is the same: the unfamiliarity disqualifies R' if and only if it interferes with common knowledge of conditional preference for conformity to R'. It may or may not. Because the expectations and preferences mentioned in the definition of convention need only be general in *sensu diviso*, it does not matter if we have no general concept of action in conformity to R'. In fact, I argued in the last section that it would be all right if we had no general concept even of action in conformity to our actual convention. The unfamiliarity would matter, however, if it led one to fail to appreciate the advantage of some action in conformity to R' when presented with a particular instance of S in which the others did conform to R'; or if it led to a failure of common knowledge that one *would* appreciate that advantage.

Again there is no disagreement with the original condition, where it applies. If the only alternatives to R are disqualified by their unfamiliarity, it would not be common knowledge that S was a coordination problem. So R would not be a convention under either condition, though it might become one whenever the members of P became acquainted with the possibility of acting in conformity to R'.

What is not conventional among narrow-minded and inflexible people, who would not know what to do if others began to behave differently, may be conventional among more adaptable people. What is not conventional may become conventional when news arrives of aliens who behave differently; or when somebody invents a new way of behaving, even a new way no one adpots. When children and the feeble-minded conform to our conventions, they may not take part in them *as* conventions, for they may lack any conditional preference for conformity to an alternative; or they may have the proper preferences, but not as an item of common knowledge. I find these corollaries of our analysis of convention neither welcome nor unwelcome. The analysis is settling questions hitherto left open.

If it seems reasonable to exclude alternatives that are too unsatisfactory or unfamiliar, as not contributing to the arbitrariness of a

convention, we have a new reason to require common knowledge. For it is by means of our common-knowledge requirement that we can exclude them without doing so *ad hoc*.

4. Degrees of Convention

We have confined our attention to perfect cases of convention, to which our definition applies without exceptions. But we cannot hope to find many perfect specimens in reality. It is time to be less strict, to allow for conventions that meet the present definition only for the most part or with high probability. Let us assemble the definition as amended so far:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in any instance of S among members of P,
>
> (1) everyone conforms to R;
> (2) everyone expects everyone else to conform to R;
> (3) everyone has approximately the same preferences regarding all possible combinations of actions;
> (4) everyone prefers that everyone conform to R, on condition that at least all but one conform to R;
> (5) everyone would prefer that everyone conform to R', on condition that at least all but one conform to R',
>
> where R' is some possible regularity in the behavior of members of P in S, such that no one in any instance of S among members of P could conform both to R' and to R.

We can count many explicit and implicit universal quantifications; we want to find a reasonable way of relaxing some or all of these to almost-universal quantifications.

The common-knowledge requirement involves universal quantifications over P (see the definition of common knowledge). We need

not allow any exception to these; anyone who might be called an exception might better be excluded from *P*. It follows, however, that most of our specifications of a population in which a convention holds will be only approximately correct.

There is no harm in allowing a few abnormal instances of *S* which violate some or all of clauses (1)–(5). So we replace "in any instance of *S* among members of *P*" by "in almost any instance of *S* among members of *P*." If we ever want more precision, we can replace it by "in a fraction of at least d_0 of all instances of *S* among member of *P*" with d_0 set slightly below one.

Nor is there any harm in allowing some, or even most, normal instances of *S* to contain a few abnormal agents who may be exceptions to the initial universal quantifications in some or all of clauses (1)-(5). So we replace each initial "everyone" by "almost everyone" or by "everyone in a fraction of at least d_i of all those involved," with each d_i set slightly below one. (We have d_1 for clause (1), d_2 for clause (2), d_3 for clause (3), and d_4 for clauses (4) and (5)—the same for both, since they are intended to be parallel.)

If we allow there to be a few agents who will not conform, we should allow the rest of the agents to know it; so "everyone else" in clause (2) should be replaced by "almost everyone else" or by "everyone else in a fraction of at least d_1 of all those involved." And if we allow the agents not to expect perfect conformity, we must not make their preferences for conformity conditional upon otherwise perfect conformity; otherwise we would not guarantee that they did prefer conformity in most cases. Their preferences should be such that if enough conform, then the more the better. (So one thing we do *not* tolerate is a convention to which most people want there to be exceptions, however few the exceptions they want.) Clause (4) should therefore be amended again to read "prefers that any one more conform to *R*, on condition that almost everyone conform to *R*" or "prefers that any one more conform to *R*, on condition that a fraction of at least d_1 of all those involved conform to *R*." Although this amendment makes clause (4) more strict rather than less, it is

unavoidable given our relaxation of clauses (1) and (2). Clause (5) should be amended in the same way to keep it parallel to (4).

We may also tolerate a few exceptions to the required incompatibility between R and its alternative R'—exceptions for most agents in a few instances of S, for a few agents in most instances of S, or both. We replace the incompatibility clause by "such that almost no one in almost any instance of S among members of P could conform both to R' and to R," or by "such that for a fraction of at least d_5 of all pairs of an instance of S among members of P and an agent therein, the agent could not conform both to R' and to R," with d_5 set slightly below one.

Our final definition is therefore:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in almost any instance of S among members of P,
>
> (1) almost everyone conforms to R;
> (2) almost everyone expects almost everyone else to conform to R;
> (3) almost everyone has approximately the same preferences regarding all possible combinations of actions;
> (4) almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R;
> (5) almost everyone would prefer that any one more conform to R', on condition that almost everyone conform to R',
>
> where R' is some possible regularity in the behavior of members of P in S, such that almost no one in almost any instance of S among members of P could conform both to R' and to R.

If anyone complains that our final definition of convention is imprecise, he is welcome to use the following quantitative definition.

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention*

to at least degrees d_0, d_1, d_2, d_3, d_4, d_5 if and only if it is true that, and it is common knowledge in P that, in a fraction of at least d_0 of all instances of S among members of P,

(1) everyone in a fraction of at least d_1 of all those involved conforms to R;
(2) everyone in a fraction of at least d_2 of all those involved expects everyone else in a fraction of at least d_1 of all those involved to conform to R;
(3) everyone in a fraction of at least d_3 of all those involved has approximately the same preferences regarding all possible combinations of actions;
(4) everyone in a fraction of at least d_4 of all those involved prefers that any one more conform to R, on condition that a fraction of at least d_1 of all those involved conform to R;
(5) everyone in a fraction of at least d_4 of all those involved would prefer that any one more conform to R', on condition that a fraction of at least d_1 of all those involved conform to R',

where R' is some possible regularity in the behavior of members of P in S, such that for a fraction of at least d_5 of all pairs of an instance of S among members of P and an agent involved therein, the agent could not conform both to R' and to R.

He may go on to define a convention as any regularity that is a convention to at least certain set degrees, which he may pick however he likes.

Let us define the *degree of conventionality* of a regularity R as the set of sextuples $\langle d_i \rangle$ such that R is a convention to at least degrees d_0, d_1, d_2, d_3, d_4, d_5. We can compare regularities with respect to their degrees of conventionality: R_1 is *more conventional* than R_2 if and only if the degree of conventionality of R_2 is a subset of the degree of conventionality of R_1. It would be interesting to find a single number that measures the degree of conventionality of a regularity; but all the ways I know to do this seem very artificial. If R is a

convention according to the strict definition at the beginning of this section, then R is a convention to at least degrees 1, 1, 1, 1, 1, 1, and no other regularity can be more conventional.

5. Consequences of Conventions

Suppose R is a conventional regularity; and suppose R^* is some logical consequence of R. Is R^* therefore a convention in its own right?

There are trivial consequences of conventions, and we are not concerned with these. Let R be our convention of driving on the right; a logical consequence of R is that we drive on the surfaces of the roads, not ten feet in the air or ten feet underground. More trivially still, a tautology that is a consequence of anything is a consequence of any convention. What we want to consider are the consequences of conventions which *depend* on convention. Our consequence R^* depends on R only if there is a regularity R' that is an alternative to R (in the sense of section 3) and *not*-R^* is a logical consequence of R'.

If so, R^* may be a convention. Suppose you and I want to meet every week; and suppose we spend alternate weeks in different towns *T1* and *T2*. Town *T1* has three acceptable meeting places: *P11*, *P12*, and *P13*. Town *T2* also has three acceptable meeting places, each analogous to the like-numbered place in *T1*: *P21*, *P22*, and *P23*. Our convention R is this: in the weeks we spend in *T1* we go to *P11*, and in the weeks we spend in *T2* we go to *P21*. A consequence R^* of R is this: in the weeks we spend in *T1* we go to *P11*. It is a dependent consequence, since *not*-R^* would be a consequence of most of the alternatives to R. R^* is certainly a convention. In the situations to which R^* applies—our weeks in *T1*—it is common knowledge among us that we conform to R^*, we expect each other to conform to R^*, and uniform conformity to R^* is a coordination equilibrium in a coordination problem. In general, a specialization of a convention is a convention. Perhaps a consequence of a convention is a conven-

tion in its own right only if it is a specialization of the original convention.

Now let us look at a dependent consequence of a convention which is not itself a convention. Suppose there is just one town with three acceptable meeting places: *P1*, *P2*, and *P3*. Suppose we want to meet; but in case we fail to meet, it is desirable that one of us should go to *P3*. Suppose our payoff matrix is as given in Figure 29, and suppose

	C1	C2	C3
R1	1 1	0 0	.6 .6
R2	0 0	1 1	.6 .6
R3	.6 .6	.6 .6	1 1

Figure 29

our convention *R* is to go to *P1*. Let *R** be the regularity of going either to *P1* or to *P2*. *R** is a consequence of *R*; and it is a dependent consequence, since *not-R** would follow from the regularity of going to *P3*, which is an alternative to *R*. We conform to *R** and expect each other to, and both of these facts are common knowledge between us. But *R** is not a convention because, I contend, it is not the case that each of us prefers to conform to *R** conditionally upon the other's conforming to *R**. Given only that you will conform to *R**, with no indication of whether you will do so by going to *P1* or by going to *P2*, I prefer to violate *R** by going to *P3*. The same is true for you with respect to me.

The case is not entirely clear, however. Consider that what we call a preference conditional upon some state of affairs *A* is almost always conditional also upon some background state of affairs *B*, which we

regard as a fixed part of the environment. To say that I prefer to drive on the right if others do is really to say that I prefer to drive on the right if others do *and* if various familiar facts about the causes and effects of collisions continue to hold. Now it is a fact, and common knowledge between us, that if either of us conforms to R^*, he will do so by conforming to R; in other words, it is a fact and common knowledge that we will not go to *P2*. If this fact were included in the fixed background, then each of us *would* prefer to conform to R^*, conditionally upon the other's conforming to R^* and upon background. I am sure it is wrong to include in the fixed background this fact, that if either of us conforms to R^* it will be by conforming to R. But I have no theory to explain why it is wrong. Roughly, the reason is this: in considering preferences for actions conditionally upon actions, the background ought to be kept neutral as to whether actions of the general sort under consideration are done or not.