The Phylogeny of Rationality

JOHN L. POLLOCK

University of Arizona

A rational agent has beliefs reflecting the state of its environment, and likes or dislikes its situation. When it finds the world not entirely to its liking, it tries to change that. We can, accordingly, evaluate a system of cognition in terms of its probable success in bringing about situations that are to the agent's liking. In doing this we are viewing practical reasoning from "the design stance." It is argued that a considerable amount of the structure of rationality can be elicited as providing the only apparent solutions to various logical and feasibility problems that arise in the course of trying to design a rational agent that satisfies this design specification.

Rationality represents one solution to the problem of survival in a hostile world. A rational agent has beliefs reflecting the state of its environment, and it likes or dislikes its situation. When it finds the world not entirely to its liking, it tries to change that. Its cognitive architecture is the mechanism whereby it chooses courses of action aimed at making the world more to its liking.

This is the general conception of rationality that I am concerned to investigate. This should be understood as a stipulation regarding what I mean here by "rationality." As I have described it, a rational agent has beliefs and likes or dislikes. At the moment, I am using these terms loosely. A rational agent must have some internal "doxastic" states that are at least fairly well correlated with some states of its environment, and some conative dispositions to "like or dislike" its situation. The agent must in addition have cognitive mechanisms whereby all of this leads it to engage in activity that has a tendency to change its immediate environment in such a way that it acquires new doxastic states that interact with its conative dispositions with the result that the agent likes its situation better.

The preceding description of rationality amounts to a crude functional description of what rationality is trying to accomplish, and how it seeks to do that. Given this description, we can evaluate a system of cognition in terms of its probable success in bringing about situations that are to the
agent’s liking. To use Dennett’s (1987) felicitous term, we are viewing cognition from “the design stance.” It will be argued that a considerable amount of the structure of rationality can be elicited as providing the only apparent solution to various logical and feasibility problems that arise in the course of trying to design a rational agent that satisfies this design specification. The argument amounts to a logical reconstruction of the phylogeny of rationality.

1. PRECURSORS OF RATIONALITY

As I will use the term, an agent is any system capable of acting on its environment to render it more congenial to its continued survival. This is a very broad notion of an agent that includes both plants and animals, and even viruses. The simplest kind of agent is a noncognitive agent. Rather than directing activity by anything approximating thought and reasoning, a noncognitive agent simply reacts. The simplest kind of noncognitive agent has built-in reflexes that respond directly to perceptual input. The behavior of such an agent is entirely hard-wired. It is incapable of learning.

A more complex kind of noncognitive agent is capable of acquiring new reflexes through something like operant conditioning. This represents a significant step upwards. For operant conditioning to be possible, there must be states of the agent-cum-environment that act as positive or negative reinforcers when they follow actions taken by the agent. Speaking loosely, these are situation types that are “liked” or “disliked” by the agent. Such conative responses to situation types are not required in a simpler noncognitive agent that is incapable of learning. They represent the first step along the road to rationality.

Noncognitive agents respond directly to perceptual input. The next step towards rationality results from adding more sophisticated mental representations of the world. Cognitive agents have both mental representations of the world and likes and dislikes. When their representations indicate that the world is not entirely to their liking, they engage in activity aimed at changing that. Epistemic cognition is the mechanism responsible for the formation of these mental representations. Once they become sufficiently sophisticated, we regard the mental representations as beliefs.

There are different solutions to the problem of generating activity in response to beliefs coupled with likes and dislikes. The simplest mechanism would consist of built-in reflexes, as in a simple noncognitive agent. A more sophisticated cognitive agent will also exhibit conditioned reflexes. The difference between conditioning in a cognitive agent and conditioning in a noncognitive agent is that in the former responses can be conditioned to arbitrary mental representations of the world and not just to perceptual input. It is worth noticing that in order for conditioning to be possible, some initial dispositions to act have to be built-in. Otherwise, the agent would
engage in no activity prior to conditioning, and hence no conditioning could occur.

In principle, there could be cognitive agents capable of sophisticated epistemic cognition but only capable of directing their activity through built-in or conditioned reflexes. I will refer to such an agent as a reactive agent. More sophisticated cognitive agents add practical cognition to epistemic cognition. Practical cognition directs activity on the basis of the agent’s beliefs about its environment, and does so in more sophisticated ways than by built-in or conditioned reflexes. Practical cognition enables the agent to direct its activity by thinking about how best to act given its beliefs about its situation and its likes and dislikes. Practical cognition can vary in complexity. A “shortsighted” agent may direct its activity one act at a time, on the basis of the observation that one act is more likely to produce a situation the agent likes than is some alternative act. A more sophisticated agent can engage in complex planning. A simple plan may require performing several acts in a fixed order. The execution of a more complicated plan might span a lifetime. (See Pollock, 1992, for an argument that choosing acts and choosing plans are genuinely different forms of practical reasoning.)

2. MECHANISMS OF COGNITION

Among the mechanisms of epistemic and practical cognition are some we recognize as reasoning. The reasoning involved in epistemic cognition has generally been called “theoretical reasoning,” but that is somewhat misleading, so I will call it epistemic reasoning. The reasoning involved in practical cognition is practical reasoning. I will refer to epistemic and practical reasoning jointly as ratiocination. Without, at this point, trying to characterize reasoning, we can nonetheless recognize that there are other cognitive mechanisms that also produce beliefs and direct activity. This is required by the fact that reasoning is slow. Many aspects of reasoning are essentially serial. Whenever possible, human cognition is accelerated by employing our inherently slow hardware in parallel processing. However, much of reasoning cannot be done in parallel, so human cognition includes many nonratiocinative processes that also issue in beliefs and actions. For instance, a human being does not have to pull out a pocket calculator and compute trajectories in order to catch a baseball or get out of the way of a bus. We have a built-in mechanism that allows us to estimate such trajectories very quickly.

Such nonratiocinative processes have sometimes been called “quick and dirty,” but that is a misnomer because they need not be at all dirty. For example, our built-in procedures for computing trajectories are incredibly accurate. Their shortcoming is not inaccuracy but inflexibility. They achieve their speed by building in assumptions about the environment, and when those assumptions fail, then the processes may yield wildly inaccurate answers.
For instance, if we see that a baseball is going to bounce off a telephone pole, we had best wait until it ricochets before predicting its trajectory. Our built-in "trajectory module" cannot handle this situation accurately, so we use ratiocination to override it temporarily until the situation becomes one that can be handled accurately by the trajectory module. I will refer to modules like the trajectory module as Q&I modules ("quick and inflexible").

It must not be supposed that human Q&I modules are concerned exclusively with motor skills. Psychological evidence strongly suggests that most everyday inductive and probabilistic inference is carried out by Q&I modules (Tversky & Kahneman, 1974). In this case, the modules really are rather dirty. Accurate probabilistic reasoning is in many cases computationally infeasible, and so humans appear to rely upon processes like Tversky's (1977) "representativeness heuristic," which often yield results incompatible with the probability calculus. This is not to say, however, that it is unreasonable to rely upon such approximation methods. The alternative of explicit reasoning is too slow for many practical purposes, and very rough approximations often suffice for the purposes for which we need the probabilities.

Thus far I have only talked about Q&I modules for belief formation, but it will emerge that Q&I modules are equally important in practical cognition. In addition to the practical Q&I modules discussed later, I suspect that in at least some cases, emotions constitute Q&I modules for practical reasoning. For instance, being afraid of tigers initiates quick avoidance responses without our having to think about it. This is a very useful thing for anyone who is apt to encounter tigers unexpectedly. Embarrassment, indignation, and so forth, may similarly be practical Q&I modules whose purpose is to supplement explicit practical reasoning in social situations. This provides a computational role for these emotions and throws light on why human beings are subject to them. This account is not applicable to all emotions. For instance, it seems unlikely that depression can be viewed as a practical Q&I module. But perhaps what this shows is that what have been called "emotional states" really constitute a grab bag of different kinds of states playing different roles in human beings.

The advantage of Q&I modules is speed. The advantage of ratiocination, on the other hand, is extreme flexibility. It seems that ratiocination can, in principle, deal with any kind of situation, but it is slow. In complicated situations we may have no applicable Q&I modules, in which case we have no choice but to undertake explicit reasoning about the situation. In other cases, human beings accept the output of the Q&I modules unless they have some explicit reason for not doing so. Ratiocination is used to monitor the output and override it when necessary. In sum, the role of ratiocination should be (1) to deal with cases to which Q&I modules do not apply, and (2) to monitor and override the output of Q&I modules as necessary. A rational agent can be viewed as a bundle of Q&I modules with ratiocination sitting on top and tweaking the output as necessary.
A cognitive agent could be devoid of ratiocination. Its cognition could consist entirely of Q&I modules. It is quite possible that many moderately sophisticated animals work in this way. I will take the defining characteristic of rationality to be the supremacy of ratiocination over Q&I modules. A fully rational agent is one in which ratiocination has total overall control. This is not to denigrate the importance of Q&I modules. It is doubtful that any agent could survive in a hostile world relying upon ratiocination as its exclusive means of cognition, but what sets rational agents apart from others is that ratiocination provides the court of last appeal. Given time to make the appeal, the agent will direct its activity in accordance with the pronouncements of ratiocination insofar as these differ from the pronouncements of its Q&I modules. We can divide rationality into epistemic rationality and practical rationality. In a fully epistemically rational agent, epistemic reasoning takes ultimate precedence over Q&I modules in belief formation, and in a fully practically rational agent, practical reasoning takes ultimate precedence over Q&I modules in directing activity. It will become apparent that human beings are closer to being fully epistemically rational than they are to being fully practically rational.

3. THE LOGICAL STRUCTURE OF PRACTICAL RATIONALITY

Situation-likings
The design objective of rational cognition is to direct activity that will make the world more to the agent's liking. What the agent likes, in this sense, is its total (token) situation-at-a-time. It likes "the way things are." I will refer to these as situation-likings. We can, accordingly, evaluate a system of cognition in terms of its probable success in bringing about situations that are to the agent's liking. In doing this we are taking an "external" or "instrumental" perspective on cognition, viewing it from the design stance. My strategy will now be to argue that a considerable amount of the structure of rationality can be elicited as providing the only apparent solutions to various logical and feasibility problems that arise in the course of trying to design a rational agent that satisfies this design specification.

Situation-likings provide the ultimate starting point for rational deliberation. These are not representational states. The agent need not be thinking about the way things are, or even know how things are. Situation-liking is a feeling rather than a propositional attitude. Instead of saying that a rational agent seeks to make its situation more to its liking, the same point could probably be made by saying that the agent seeks to make itself happier, or seeks pleasure. I have avoided these more traditional formulations because the terms "happiness" and "pleasure" have connotations not intended by the theory. Situation-liking is perhaps best described as a "pro-attitude" towards one's current situation. This is only explanatory insofar as it is
understood as pointing towards a fuller functional explanation of situation-likings, and that is the way I think they ultimately should be understood. The claim is simply that there is a mental state, here called "situation-liking," that plays a certain functional role in practical cognition, and the rest of the theory of practical cognition fills out the description of that functional role. It will emerge later that, as human beings, we have introspective access to situation-likings. This gives us a grasp of them that is largely independent of their functional role, and makes them more familiar to us than if we had to simply postulate them as "unobserved states that play a certain role in cognition."

Plans
As I shall use the term, to live the good life is to live so that, in so far as possible, one likes the situation one is in. Practical cognition seeks to direct the agent in its pursuit of the good life. It does this by selecting and executing courses of action that are likely to contribute to this objective. Such a course of action is a plan.

Planning aims at the construction or discovery of plans whose execution is likely to put the agent in a situation it likes better than the one it would otherwise be in. An agent may be faced with a choice between several plans, each of which is likely to improve the agent's situation. In choosing among them, the agent must balance their likelihood of improving its situation against the degree of improvement they make likely. More generally, in evaluating a plan, an agent should ideally consider each possible outcome and how much it would like that, and weight the outcome proportional to the agent's estimate of the probability of that being the actual outcome. To use standard mathematical terminology, plans should be evaluated in terms of the mathematical expectation of the agent's liking of the situation resulting from adopting the plan. In thus evaluating a plan, we are never in a position to predict what (infinitely specific) situation token will result from its execution. All that we can predict is that, with various probabilities, certain types of situations will result, the types being defined in terms of their features. We evaluate the plan by evaluating those different situation types and forming a weighted average, weighting each according to our estimate of the probability that it will be the situation type resulting from executing the plan.

Expected Situation-likings
To evaluate plans in this way, we must first be able to evaluate situation types. This leads to a kind of type/token problem. The likes and dislikes that define the goal of practical reasoning pertain to situation tokens, but the reasoning itself can only proceed in terms of situation types. Such reasoning is made possible by the fact that our liking or disliking a situation token is caused by general features of the situation. For example, after a strenuous
hike into a beautiful mountain valley, being hungry and tired disposes me towards disliking the current situation, but being in a spectacular place disposes me towards liking it. The only way a rational agent can pursue liked situation tokens and seek to avoid disliked situation tokens is by identifying features causally relevant to their being liked or disliked and then pursuing or avoiding those features. The objective of practical reasoning is to render the actual situation token likeable (or as likeable as possible), but the only way to do this is to try to ensure that it has features that tend to render situation tokens likeable. (Pettit, 1991, made similar observations, although he identified likes with desires.) Accordingly, a situation type must be evaluated in terms of its "expected situation-liking," that is, the mathematical expectation of the degree of liking of situation tokens of that type.

The expected liking of a situation type is a weighted average of the likings of situation tokens of that type, where the weighting is determined by the probability of tokens of the type being liked to any particular degree. If we make the simplifying assumption that there are just finitely many possible values $r_1, \ldots, r_n$ for the situation-liking of tokens of type $F$, then the expected liking of tokens of type $F$ is the sum

$$\sum_{i \leq n} r_i \cdot \text{prob}(\text{the situation-like of } x \text{ is } r_i / x \text{ is a situation token of type } F)$$

In the general case the expected-liking is

$$\int_{-\infty}^{+\infty} r \cdot d\text{prob}(\text{the situation-liking of } x \text{ is } r / x \text{ is a situation token of type } F).$$

Only by employing such evaluations can practical reasoning aim at rendering the actual situation token likeable.

The general goal of rationality is to maximize the agent's situation-liking. For this to be well defined, all we need is a comparative relation of liking one situation token more than another. But now we find that the only way a rational agent can attempt to improve its situation-liking is by reasoning about situation types in terms of their expected situation-likings. This introduces a serious complication. In order for the expected situation-liking of a situation type to be well defined, it is no longer sufficient to have a comparative relation of situation-liking: It must be possible to assign numerical values to the degree of situation-liking an agent has for its current situation. That there are such numbers does not seem particularly problematic, because the degree of situation-liking is just a physical parameter of some sort that plays a certain role in the agent's cognitive architecture. However, there remains a problem of "scale." Mathematical expectations require not just numerical measures, but cardinal measures. The result of adding or averaging such measures must make sense. This is not guaranteed by the mere existence of a measure. For example, given one measure of the degree of situation-likings, the logarithm of that measure is also a measure, but adding one of
the measures is equivalent to multiplying the other. This has the consequence that these two measures can yield different comparisons of expected situation-likings, so that on the first measure a situation type $S_1$ may have a higher expected situation-liking than a second situation type $S_2$, whereas on the second measure $S_2$ will have a higher expected situation-liking than $S_1$. In the standard theory of measurement, the problem of choosing a scale is solved by supposing there is a physical composition process that corresponds to applying addition to the measurements, but no such process presents itself in the case of situation-likings. How, then, are we to choose a particular measurement scale for situation-likings?

The key to this first problem is to realize that there is a second problem that must also be solved. If a rational agent is to employ expected situation-likings in its rational deliberations, it must be able to form beliefs about their values. This means that it is not enough for there to merely exist a numerical measure of situation-likings. The agent must also be able to form beliefs about the values of this measure. The only obvious way to do this is through introspection. An ideally designed rational agent would be able to introspect its situation-likings, and the introspection would yield a number that could then be employed in computing expected situation-likings. Such introspection would simultaneously solve the scale problem and the problem of forming beliefs about situation-likings. In effect, the proper scale becomes defined functionally as the scale used by introspection, and the results of introspection are then employed in various ways in cognition.

It must be emphasized, however, that the preceding remarks are about introspection in an "ideally designed" rational agent. Two observations are in order about introspection. First, introspection tends to be in disfavor these days. This is partly a legacy of logical positivism, and partly a result of psychological experiments that show introspection to be unreliable as a guide to certain aspects of cognition (see Lackner & Garrett, 1972). In the latter connection, it should be emphasized that the psychological investigations focus primarily on our ability to introspect causes. It is hardly surprising that we are unable to introspect causes, because causes are not the sort of thing that can be perceived: They must be inferred from simpler facts that can be perceived. The investigations actually presupposed that the subjects could tell introspectively what they were thinking, because in order to determine what their subjects were thinking, the investigators simply asked them. I have argued (Pollock, 1989) that introspection plays an essential functional role in epistemic cognition, and now I am arguing that it plays an equally essential functional role in practical cognition. As much as one's philosophical prejudices might incline one to eschew introspection, it cannot be done. Introspection is an essential and ineliminable part of cognition.

From a computational point of view, there is nothing at all mysterious about introspection. For instance, in designing a system of automated reasoning that runs on a digital computer, if a certain numerical parameter is
used to guide some bit of reasoning and it is also desirable for the reasoner
to have a belief about the value of the parameter, it is computationally trivial
to have the reasoning program determine the value of the parameter and insert
a belief about it into the reasoner's set of beliefs. That is all that introspec-
tion need amount to. There is, accordingly, no obstacle (at this point) to
constructing a rational agent capable of the kind of introspection of situa-
tion-likings that was previously described. But at the same time it must be
acknowledged that such introspection is problematic for human beings. We
are certainly capable of telling introspectively whether we like our situation,
but introspection does not provide us with a number. I suspect that this
reflects, in part, the fact that computation in human beings works quite dif-
f erently from computation in digital computers, and this makes the in-
trospection of numerical values more difficult. However, a numerical
measure of situation-liking must somehow be forthcoming if we are to be
able to compute expected situation-likings.

One of the impressive accomplishments of "rational choice theory" is
that the methods of Ramsey (1926/1960) and von Neumann and Morgenstem
(1944) provide a way of obtaining numerical measures of preference from
nonnumerical preference and indifference orderings, and it might be sup-
posed that this provides the solution to the problem of obtaining a cardinal
measure of situation-liking. However, this work is not applicable to the pre-
 sent problem. It requires preferences between situation types (specifically,
situation types consisting of gambles), but the preferences generated by situ-
atation-likings are between situation tokens.

There is a solution to this problem. It lies in the fact that although intro-
spection does not provide us with a number directly, it does provide us with
more than a preference ordering. In a rough sense introspection also apprises
us of how much we like or dislike our situation. We might more accurately
(but somewhat obscurely) say that introspection provides us with a "quan-
titative feel." It is quantitative in the sense that it not only generates com-
parisons—we like one situation better than another—it also makes sense to
ascribe rough magnitudes to the comparisons: We may like situation A a lot
better than situation B, while liking situation C only a little better than situ-
atation D. If these quantitative feels are to be used in explicit reasoning about
expected situation-likings, they must somehow be transformed into numbers.
It appears that there is, in fact, a way of doing this, as I will now explain.

We can define a quaternary preference relation $\rho(A, B, C, D)$ as holding
iff the agent would prefer having $B$ to having $A$ more than it would prefer
having $D$ to having $C$. The observation that introspection provides us with a
quantitative feel can now be expressed less obscurely as the observation that
it informs us of this quaternary relation on situation tokens. A binary pref-
erence can be defined in terms of $\rho$ by saying that $B$ is preferred to $A$ iff
$\rho(A, B, A, A)$ holds. An ordinal measure of situation-liking is simply a
function $u(x)$ such that for any $x$ and $y$, $x$ is preferred to $y$ iff $u(x) > u(y)$. If
we make reasonable assumptions about binary preference, this ensures that infinitely many ordinal measures of situation-liking exist. The scale problem is then the problem of choosing a "privileged" ordinal measure to use in computing expected situation-likings. This can be done by requiring an appropriate measure $u(x)$ to also represent the quaternary preference relation, in the sense that for any $x, y, z, w$, $\rho(x, y, z, w)$ holds iff $u(y) - u(x) > u(w) - u(z)$. Nonlinear transformations of measures will have the effect of reversing quaternary preference relations provided there are "enough" items in the domain of the preference relation, so if we make some reasonable assumptions about the set of possible situation tokens, the result will be that $u(x)$ is uniquely determined up to linear transformation. That is sufficient to yield unique comparisons of expected situation-likings.

Exactly what assumptions should be made about the set of possible situation tokens is a matter for further investigation. Different combinations of assumptions will yield the desired result. One triple of assumptions that, when added to the ordering assumptions, appears to be sufficient is as follows, where "$<$" symbolizes binary preference between situation tokens:

\[
\text{If } \gamma \text{ and } \lambda \text{ are sets of situation tokens such that for every } x \in \gamma \text{ and } y \in \lambda, \ x < y, \text{ then there is a } z \text{ such that for every } x \in \gamma \text{ and } y \in \lambda, \ x < z < y.
\]

\[
\text{For any situation tokens } x, y, z \text{ such that } x < y \text{ and } x < z, \text{ there are situation tokens } w, v \text{ such that } z < w < v \text{ and } \sim \rho(x, y, w, v) \text{ and } \sim \rho(x, y, v, w).
\]

\[
\text{For any situation tokens } x, y, z \text{ such that } x < y \text{ and } z < y, \text{ there are situation tokens } w, v \text{ such that } w < v < z \text{ and } \sim \rho(x, y, w, v) \text{ and } \sim \rho(x, y, v, w).
\]

However, there may be other sets of assumptions that would be preferable for various reasons.

The preceding observations have the consequence that expected situation-likings are well defined. They also tell us how a rational agent might discover their values through reasoning. However, it is obvious that this is not the way practical cognition normally proceeds in human beings. What seems to happen in humans is that the quantitative feels produced by introspection of quaternary preference are combined nonratiocinatively with beliefs about probabilities to produce similar quantitative feels reflecting expected situation-likings. This involves a kind of "intuitive integration." This represents a Q&I module that, I suspect, plays a pervasive role in human cognition. It enables humans to circumvent what would otherwise be some very difficult problems of epistemic reasoning. In an agent like a human being in which introspection of situation-liking does not yield numbers, the availability of such a Q&I module is extremely important. It would be of lesser importance in an agent in which introspection were more informative, but it would still be valuable.
The Identification Problem and Feature-likings

Given a solution to the logical problem of assigning numerical values to situation-likings and thereby rendering expected situation-likings well defined, a rational agent who wants to use these numbers in practical reasoning still faces the problem of identifying features or combinations of features that are causally relevant to situations being liked or disliked and determining their degree of causal relevance, that is, evaluating their expected situation-likings. I will refer to this as the identification problem. Let us call these causally relevant features “value-laden features.” As I have described it, the identification problem is, in principle, an empirical problem to be solved by epistemic reasoning aimed at discovering how such features and their combinations affect our likings for situation tokens. An ideal rational agent, unconstrained by time or resource limitations, and able to survive in the real world for an extended period of time without engaging in any practical reasoning, could solve this problem in this way and use the resulting solution to drive its subsequent practical reasoning. However, human beings are not ideally situated in the world and do not have the luxury of waiting until they have solved such complex empirical problems before they can begin engaging in practical reasoning for the first time. Nor, most likely, will any rational agent have that luxury. For this reason, human beings do not work in the way described, and it is probably impossible for any real agent to work in this way. This is a feasibility problem. Even if an ideally rational agent could solve this problem by empirical investigation of the causal structure of the world, real agents must take shortcuts vitiating the need to acquire all this theoretical knowledge prior to engaging in practical reasoning. By definition, any such shortcut procedure constitutes a Q&I module.

Exactly how human beings solve this problem is an empirical question requiring psychological investigation. Armchair philosophy cannot provide a definite answer, but introspection suggests a mechanism. Human beings have the ability to imagine a situation and then have conative responses to the imagined situation. For example, imagining a situation in which I am being torn apart on a torture rack, I exhibit an introspectible dislike for the imagined situation. It is a brute fact about human beings that they have such conative reactions to imagined situations. Their ability to react conatively to imagined situations is essential to their enjoyment of literature, and I have even suggested elsewhere that moral judgments may be based in part on this ability (Pollock, 1986a). It is important to recognize that imagined situations are always inherently general. Imagined situations can never incorporate

---

1 This may be overly optimistic. Empirical investigation is typically driven by practical reasoning regarding how to find the answers to certain questions. Accordingly, it might be impossible for even an ideal rational agent to acquire the requisite knowledge of expected likings before engaging in practical cognition.
more than finitely many features. As such, our reaction to these imagined situations constitutes a conative response to situation types rather than situation tokens. We might call these feature-likings, in contrast to situation-likings, to emphasize that they are conative responses to types rather than tokens, although it is not clear that these two kinds of likings should be regarded as genuinely different kinds of mental states. One can conjecture that the way feature-likings work neurologically is by marshaling the ordinary mechanisms that give rise to situation-likings and applying them to imagined situations, the result being an introspectible liking for an imagined situation type rather than a situation token. Human beings employ these introspectible feature-likings to make intuitive assessment of how much they would like being in situations of the type imagined, that is, to assess expected situation-likings of situation types. There is, however, no logical connection between such introspectible likings for imagined situations and the expected situation-liking of the type of situation imagined, so the former cannot provide a logically good reason for the latter. Accordingly, this kind of assessment must be regarded as constituting a Q&I module.

It is a consequence of my stipulative definition of rationality that situation-likings are the basic evaluative states in terms of which plans are to be assessed. However, we could explore an alternative conception of rationality that begins with feature likings. On this conception a rational agent pursues liked features and seeks to avoid disliked features. This approach must provide a solution to several problems. First, we need a mechanism for combining the likings of individual features to compute a liking for combinations of features. We cannot just sum the individual feature-likings, because features may interact. For instance, I like dill pickles and I like vanilla ice cream, but I do not care much for the combination. Second, a distinction must be made between the "primitive" liking for a combination of features, and the "all-things-considered" liking. The former just measures how much the agent likes those features by themselves, independent of their tendencies to be accompanied by, or to cause, other features. The all-things-considered liking takes such tendencies into account. Notice that plans must be evaluated in terms of the all-things-considered liking for adopting the plan. Typically, the agent's primitive liking for adopting a plan is neutral, and value only attaches to the plan indirectly by way of tendencies that has to bring about other features the agent likes or dislikes. It looks like these all-things-considered likings must be computed in terms of expected values, which in turn requires a cardinal measure for primitive feature-likings. The logical, mathematical, and epistemological problems faced by this approach are no less onerous than those that arise from taking situation-likings to be basic.

I see no reason in principle why an agent could not be made to conform to either conception of rationality. The overall structure of the agent would be the same on either conception, but somewhat different mechanisms would
be used for evaluating plans. Which sort of agent is a human being? There is reason to believe that the norms of rationality incorporated into the functional description of human cognition give precedence to situation-likings over feature-likings. If some feature seems desirable to a person in the abstract, but getting it invariably makes the person unhappy (and not because of concomitant features that the person dislikes), then we would regard it as irrational for the person to continue pursuing that feature. This suggests that human beings conform to the first conception of rationality, according to which feature-likings provide the court of last appeal. But the fact that human beings work in this way is not, as far as I can see, a reason for thinking this is a "better" conception of rationality.

**Plans and Goals**

Recall that "to live the good life" is to live so that, in so far as possible, one likes the situation tokens one as in. Rationality is a tool for living the good life. A rational agent directs its activity on the basis of its beliefs about the expected values of combinations of features, trying always to better its situation, that is, to render it more to its liking. It does this by choosing goals whose achievement will have that effect, and then selecting and executing courses of action that aim at the achievement of those goals. For the sake of logical uniformity, we can identify goals with combinations of features of situations. Note that these goals need not be feature combinations that the agent values highly. A rational agent may pursue things that it actively dislikes, provided it likes them better than what it already has. Similarly, a rational agent may decline to pursue things it likes, because it likes what it already has better.

An agent tries to achieve goals by designing and executing courses of action aimed at realizing them. Designing such a course of action is planning. The simplest planning may involve selecting a single act to perform in order to achieve some readily attainable goal. More complex plans may involve multiple steps, each prescribing an act subject to conditions that may depend upon both the current situation and the outcome of the previous steps of the plan. For example, a plan for driving across town might involve taking one route if the traffic is light and another route if it is heavy. It can be argued that in their most general form, plans can at least have all the logical structure of parallel computer programs.

This commonsensical description of planning involves two distinct elements: goals and plans. The relationship between goals and plans is more complicated than might initially be supposed. Goals are chosen on the basis of their expected situation-likings, but the mere fact that a plan successfully achieves such a goal is not sufficient to make it a good plan. This is because the situation token that results from executing the plan must invariably exemplify more features than those included in the combination of features
constituting the goal, and some of these may reverse the agent’s liking for
the situation. For example, it may be that, in the abstract, I would like being
rich. But it may also be that the only way I could become rich under the
actual circumstances, is by doing things I would despise doing, and I would
not like a situation that includes my being rich as a result of my doing those
things.

Goals are chosen before plans are adopted for achieving them, so goals
can only be selected on the basis of expected situation-likings in the abstract.
Accordingly, the expected value of a play may not be a function of the
abstract values of the goals it aims at achieving. This can leave one wondering
of what use are goals. Why not dispense with goals altogether? This suggests
that what a rational agent should do is just survey possible plans and select
those having high expected values. Goals would play no role in this process.

However, this suggestion runs afoul of feasibility considerations. The
construction and evaluation of plans is a difficult and time-consuming pro-
cess. An ideal rational agent, unconstrained by time or resource limitations,
could systematically generate all possible plans and select for execution those
having high expected values. However, agents subject to realistic limitations
must focus their attention more efficiently on plans that are apt to be relevant
to their current situation. They do not have the luxury of being able to pick
systematically through an infinite list of plans, most of which are irrelevant
and of no value anyway.

It is useful to compare planning with automated theorem proving. It has
often been observed that it would be possible to build an automated theorem
prover that systematically generates all possible proofs in the predicate cal-
culus. Then if we want to know whether some particular formula is a theorem
of the predicate calculus, we just wait and see whether a proof of it is gener-
ated. This is the so-called “British Museum algorithm.” Such a theorem
prover would be hopelessly inefficient for any practical purposes. By con-
trast, human reasoning is interest driven in the sense that the course of our
reasoning is determined in part by what questions we are trying to answer.
If we want to know what time it is in Moscow, we reason quite differently
than if we want to know the sum of 14 + 87. In other words, our reasoning
involves a control structure that makes it more directly germane to the ques-
tions we want to answer. (For an extended discussion of interest-driven
reasoning, see Pollock, 1990.)

A similar control structure is required for practical reasoning. A scattergun
approach to planning is as impractical as the British Museum algorithm.
Goals focus planning in two ways. First, if a plan is able to attain a goal
having an expected situation-liking higher than the agent’s expected liking
of the situation that would otherwise result, that creates a presumption in
favor of the plan having a positive expected value, because the outcome of
the plan will be a situation of a general type that tends to be better liked
than situations of the type that would otherwise result. Thus, by directing planning at such goals, the rational agent is guaranteed to focus its attention on plans that *tend* to be good plans. Second, whether a particular plan is apt to achieve a specific goal is a factual question of the sort that falls within the purview of epistemic cognition. Accordingly, the construction of plans aimed at specific goals is an exercise of epistemic cognition. As such, it can be directed by the general control structures governing interest-driven epistemic reasoning, and need not proceed at random. This would not be possible unless planning aimed at specific goals.

To summarize, feasibility constraints require the use of goals to direct the course of planning. Goals can only be evaluated in terms of expected situation-likings and without reference to any way of achieving the goals. This guarantees that the plans that achieve goals having high expected values will tend to be good plans, in the sense of producing situation tokens the rational agent likes. However, once produced, plans must be evaluated not just in terms of the values of their goals, but in terms of all predictable features of their outcomes. In other words, the expected values of plans must be computed in terms of the difference between the expected situation-liking of the situation type that consists of adopting the plan and the expected situation-liking of the situation type that consists of not adopting the plan.

**Goals and Desires**

Adopting a goal initiates planning for its achievement. It does this by passing interest in discovering such plans to the system of epistemic cognition. When an agent selects a goal, it comes to *want* it or *desire* it. Desires constitute a functionally (and, in humans, introspectively) different state from likings or valuing. Their purpose is to encode goals. Their functional role is to move the agent to act or plan, or more generally, to initiate practical cognition.

Goals are adopted as a result of being judged suitable. The suitability of a goal is a function of the agent's current situation, because goals are chosen in an attempt to better that situation, that is, to render it more to the agent's liking. It is to be emphasized that goals are not the same thing as feature combinations believed to have high expected situation-likings. As indicated previously, an agent's goal may be to achieve something it dislikes as long as it likes it better than what would happen otherwise. Suitability is a matter of *relative* expected value rather than absolute expected value.

A rational agent should adopt a goal whenever it is judged suitable in the preceding sense, but this does not guarantee that anything will actually happen as a result of adopting the goal. This is because the agent may not have any information at its disposal that will enable it to make progress in planning for the achievement of the goal. Typically, progress in planning only results when the agent recognizes an opportunity to improve its lot. For example, I like eating ice cream, but I do not spend all my days plotting
how to find and eat some ice cream. As a general rule, I only engage in planning for the eating of ice cream when I encounter a concrete opportunity for doing so. For instance, finding ice cream available as dessert at a dinner party may cause me to fabricate the plan to say "Yes" when my host asks whether I would like some. Of course, sometimes we create our own opportunities. If there is no ice cream readily available, and I want ice cream badly, I might make my own, but under ordinary circumstances, a plan to achieve the eating of ice cream in this way would not be a good plan because the effort involved would outweigh the benefits achieved.

On the preceding account, goals are adopted (desires are formed) as a result of beliefs about their expected values, and these beliefs are produced by epistemic cognition. There is a problem here, however. How does the agent know which combinations of features to evaluate as potential goals? If features are just selected randomly for evaluation, very few are going to turn out to be suitable goals. An agent must have a more efficient mechanism for proposing potential goals. The agent can then go on to evaluate these potential goals and decide which to adopt. This amounts to a control structure for that part of an agent's reasoning that is devoted to the evaluation of potential goals.

There is a second problem, analogous to some we have already discussed. A real agent cannot wait until it has acquired a lot of knowledge about the causal structure of the world before it begins to adopt goals and engage in practical cognition. Accordingly, a real agent must incorporate a Q&I module for desire formation that gets practical cognition started without the need for large amounts of theoretical knowledge.

These two problems can be solved simultaneously by (a) having a Q&I module that proposes potential goals and leads to their default adoption, and (b) giving ratiocination the power to evaluate these goals explicitly and override their adoption if they are judged unsuitable. This amounts to having overridable dispositions to form desires. I will refer to these as optative dispositions. In order to enable desire formation in situations of minimal knowledge, some of these optative dispositions must be built-in. Others can then be acquired through conditioning. For example, we have a built-in optative disposition to try to alleviate our hunger. Similarly, we have built-in optative dispositions to try to avoid pain and to pursue pleasure. We have a number of such "hedonistic" dispositions. These optative dispositions give rise to desires automatically, without our having to engage in any reasoning about their expected values. Conditioning can then lead to the establishment of new optative dispositions. When a combination of features tends to be followed by the satisfaction of desires, a conditioning mechanism can use that to create a new disposition. Thus we may come to desire money, or we may acquire the disposition to desire to read an interesting book when we hear about it, or the disposition to desire to see certain kinds of new films.
The general mechanism of built-in and conditioned optative dispositions constitutes a Q&I module for desire formation. In a rational agent, there must also be a purely ratiocinative basis for desire formation. The sole ratiocinative basis for desiring something should be the belief that it is a suitable goal. In a fully rational agent, ratiocination must have the power to override any Q&I modules. Accordingly, if an optative disposition produces a desire in me, but I have the belief that I would not like having the desire satisfied (i.e., its relative expected value is negative), this should dispel the desire. In human beings, this does not always happen, as is illustrated by compulsions. For example, a compulsive smoker may believe that the expected value of smoking is negative, but desire it nonetheless. In a fully rational agent, that belief would cancel the desire to smoke. Obviously, human beings are less than fully rational in this respect. Such beliefs may serve to depress the strength of the desire, but they do not dispel the desire entirely.

Conversely, if a fully rational agent believes that some combination of features has a high relative expected situation-liking, that should automatically produce a desire. Failure to form the requisite desire is a mark of irrationality, but people are sometimes irrational in precisely this way. For example, I enjoy skiing, but for some reason I can rarely get myself to go skiing. I know that if I went, I would enjoy it, and I am often in a position where I can take a day off and go skiing if I so choose, but I just cannot get myself motivated to do it. In other words, despite having all of the requisite beliefs, I lack the desire. This is irrational, but perhaps not terribly unusual. Once again, human beings are not fully rational agents.

**Intentions, Plans, and Instrumental Desires**

Desires provide the trigger for planning. Constructing a plan is a matter of epistemic cognition. Whether a plan will, with a certain probability, achieve some goal, is a factual question of the sort that falls within the purview of epistemic cognition. So is the evaluation of the plan in terms of expected situation-likings. However, once all that has been done, the agent still has to decide what plan to adopt. A satisfactory plan must be at least minimally good, in the sense that the expected situation-liking of the situation type consisting of adopting the plan is at least as great as the expected situation-liking of the situation type consisting of not adopting the plan. Equivalently, the plan must have a non-negative expected value. However, this is not automatically sufficient to make it rational for an agent to adopt the plan, because sometimes the agent will have to choose among several competing plans, and one of the competitors may be better than the others. In that case, the optimal competitor should be adopted. The logic of such choices is more complicated than it may seem at first, and it constitutes the subject matter of Pollock (1992). When, as a result of such deliberation, a plan is adopted, this must be encoded in the agent somehow. The states that do this
are intentions. The functional role of intentions is to encode plan adoption. (This observation was perhaps first made by Bratman, 1987.)

A plan is typically a "partial" solution to the problem of achieving a goal. (This is a familiar observation in AI planning theory; it is due originally to Sacerdotti, 1975.) For instance, given the goal of being in Los Angeles on Friday, I may plan to fly on Friday morning. However, in order to execute this plan I may have to engage in further practical reasoning. I will have to make plans about when and where to buy a ticket, what airline to use, what flight to take, and so on. When an agent adopts a partial plan some of whose steps require further planning, there must be something that initiates that further planning. The state having this functional role is what we call an instrumental desire. It is a desire because its role is to initiate further planning, but it is not a desire produced by any of the mechanisms so far discussed. That is, it is neither the result of optative dispositions nor of beliefs about the relative values of potential goals. Let us call desires produced in either of those ways primitive desires. Primitive desires provide the starting points for planning. In contrast to primitive desires, instrumental desires are produced by practical reasoning itself. Whenever a plan is adopted, rationality dictates that the agent form the desire to execute the plan steps at the appropriate time, and these desires in turn act either as triggers for action if the plan step is one that can be executed directly or as triggers for further practical reasoning.

Action Initiation
Having adopted some plans, there must still be something further that gets an agent to act. There is a temptation to think of the adoption of a plan as beginning a process that grinds inexorably to the execution of the plan unless the plan is subsequently retracted (i.e., unless we change our minds). But things don't really work that way. Adopting a plan is not like setting a train moving on rigid rails. Even though I plan to do something and never explicitly change my mind, the plan may never be executed just because I never feel like executing it. There may always be other things I want to do more. This can include doing nothing. For example, I might adopt the plan to go to the grocery store this afternoon. As the afternoon progresses, it occurs to me at various times that I planned to go to the grocery store, but each time my reaction is, "Oh, I don't feel like going just now. I will do it later." Finally, the afternoon is over and I have not gone, but there was never a point at which I explicitly retracted my plan to go. I just didn't do it. Furthermore, it is not only lethargy than can keep plans from being executed. I may have planned to go to the grocery store, but I became engrossed in writing this article, and every time I thought about going to the grocery store I decided that I would rather keep writing and go later. Eventually, the afternoon was over and I had not gone.
The possibility of such procrastination arises from the fact that planning typically leaves the scheduling of plan steps rather indefinite. The advantage of this is that it minimizes conflicts between adopted plans and unforeseen opportunities that arise later. But the disadvantage is that some further mechanism is required for deciding when to execute a plan step.

The basis for such a mechanism lies in a distinction between two kinds of primitive desires: desires for the future and desires for the present. Only the former can be the subject of planning. You cannot plan for the present. Planning takes time and you cannot plan ahead for a situation that is already here. In order to play any role in directing action, desires for the present must be desires to do something. I will refer to such desires as "present-tense action desires."

Recall that a reactive agent guides its actions solely on the basis of built-in and conditioned reflexes generating immediate action. However, there is a problem regarding reactive agents that I have so far ignored. This is that more than one reflex can be triggered at the same time, and it may be impossible to perform the actions dictated by them all. A reactive agent must have some mechanism for adjudicating disputes in such cases. The only obvious mechanism is to assign strengths to reflexes, and when two reflexes compete, the stronger wins. We can, at least metaphorically, think of such reflexes as generating present-tense action desires, and in deciding what to do when there is a conflict, all the reactive agent can do is act on the desire that is strongest, that is, do what it most wants to do.

An agent capable of planning can do better. It can, in many cases, adjust its plans to avoid such conflicts. However, planning can never entirely replace reacting to present-tense action desires. Planning is only applicable insofar as we can predict what is apt to befall us. If the unexpected happens, it may be too late to plan for it and all we can do is react. Basic reflexes, like withdrawing from pain, are going to be essential ingredients in any rational agent. Ratiocination can only supplement such prerational mechanisms.

Just as a reactive agent can experience conflicting reflexes, a rational agent can encounter conflicts between plans and reflexes. The plan to retrieve a valuable object that fell into a fire may conflict with the reflex of withdrawing one's hand from the fire. Sometimes, the plan wins out. This requires that there be an adjudication mechanism in a rational agent that can choose not only between actions prescribed by present-tense action desires, but also actions prescribed by plans. This must be done in terms of the value of some parameter of the actions among which the agent is choosing. In the case of actions prescribed by present-tense action desires, the relevant parameter is how much the agent wants to perform the action, that is, the strength of the desire. Accordingly, we can use the same language in talking about the parameter attaching to actions prescribed by plans. The agent wants, more or less strongly, to perform such an action, and it can
decide which action to perform by selecting the one it wants most strongly to perform. Jean Hampton pointed out to me that this is essentially a Hobbesian view of deliberation (see *The Leviathan*, chap. 6, paragraph 29). How badly an agent wants to perform such an action can be regarded as the strength of the agent's desire to perform it, where this desire is the instrumental desire produced by adopting the plan.

The degree to which an agent wants to perform an action prescribed by a primitive present-tense action desire is determined by the optative disposition producing that desire. There must be a mechanism determining the degree to which an agent wants to perform an action prescribed by a plan. Ideally, this should result in choices being made so that the plans having higher expected values are the ones that will be executed. The way to do this is to have the strength of an instrumental desire correspond to the expected value of the plan. More accurately, when a plan prescribes a sequence of acts and $A$ is one of those acts, let the *tail* of the plan relative to $A$ be the remainder of the plan after the steps preceding $A$ have been executed. The tail of a plan is itself a plan, and we can talk about its expected value in the same way we talk about the expected value of any plan. Once the first part of a plan has been executed, the degree to which an agent wants to execute the next step should correspond to the expected value of the tail of the plan relative to that next step.

Given this understanding of degrees of wanting to perform actions, my proposal is that a rational agent should, at any given time, perform the action it most wants to perform. This amounts to marshaling the prerational machinery of the reactive agent and extending it to the results of planning. However, there are what appear initially to be counterexamples to this account. It seems that it is sometimes rational to postpone doing what I most want to do if I know that I can do it later and there is something else I want to do that cannot be postponed. To take a concrete example, suppose I am out of the house, and I plan to telephone a friend when I get back and before she leaves for work. There will be only a short interval in which I can do that. As I am driving home, I become increasingly hungry and want to have lunch. If I had to choose between (a) eating lunch and not calling my friend, and (b) calling my friend and not eating lunch, I would unhesitatingly choose lunch. So it seems that I want to eat lunch more than I want to call my friend. But what I may actually do is postpone lunch just long enough to call my friend, because I know that I will not be able to call her later. This appears to be a counterexample to the account just proposed for the rational initiation of action.

However, this putative counterexample confuses the strength of the instrumental desire to eat lunch (my present-tense action desire) with the strength of my primitive desire to eat lunch. It is quite true that my primitive desire to eat lunch is stronger than my primitive desire to telephone my
friend. That is what is shown by the observation that if I had to choose between them, I would choose lunch. However, in the technical sense defined before, the strength of my wanting to eat lunch and the strength of my wanting to telephone my friend are not determined by the strengths of the primitive desires, but rather by the expected values of the plans in which they are embedded. What actually happens is that I begin by adopting the plan to telephone my friend when I get home. Then I get hungry. I consider two plans. The first is to eat lunch immediately upon returning home. The second is to telephone my friend immediately upon returning home, and then eat lunch. The second plan is preferable to the first, so it is adopted. As the execution of that plan unfolds, telephoning my friend is the first action mandated. It thus becomes a candidate for performance. At that point, eating lunch has not even been prescribed by an adopted plan. That step will not be prescribed until after I have telephoned my friend. Eating lunch is only prescribed by a primitive desire to alleviate my hunger forthwith, but that desire (we can suppose) is weaker than my instrumental desire to call my friend, because the latter derives from a plan for both eating lunch and calling my friend. Accordingly, I will first choose to telephone my friend, and then to have lunch.

4. AN OVERVIEW OF PRACTICAL RATIONALITY

The preceding discussion shows that approaching rationality from the design stance suffices to generate a great deal of the structure governing how a rational agent must function. The account has been rather complex, so let me provide a brief recapitulation.

- The objective of rationality is to live so that, insofar as possible, one likes the situations one is in. In other words, maximize situation-liking. Thus, situation-likings become the fundamental states in terms of which rationality is to be understood.
- A rational agent tries to accomplish this objective by discovering courses of action (plans) whose execution will result in its liking its situation better. Adopted plans are stored as intentions.
- In discovering such plans, the agent must reason about situations in terms of general features of them. The prediction that the execution of a plan will result in a situation the agent likes must be based on general beliefs about the expected likability of general types of situations. In particular, the agent must assess plans in terms of beliefs about the expected situation-liking of the situation type consisting of adopting the plan.
- Evaluating the expected situation-likings of situation types is an extremely difficult problem. Any agent subject to realistic constraints
must use Q&I modules to help it solve this problem. At least in human beings, these generate feature-likings, which are used nonratiocinatively to produce estimates of expected situation-likings.

- In searching for plans having high expected values, some mechanism is required to focus the search on plans that are antecedently likely to have high expected values. This is done by selecting as goals highly valued situation types and then searching for plans to achieve these goals. The search can then be constrained by using familiar techniques like goal reduction. Adopted goals are encoded as desires.

- Rational action is not just a matter of mechanically executing adopted plans. The agent must decide when to execute a plan, and must choose among competing actions available at any given time. A rational agent does this by forming present-tense action desires, which are desires to perform currently available actions. Some of these desires are produced by optative dispositions, and others are derivative from adopted plans. Rationality then dictates performing the action prescribed by the strongest such desire.

Philosophical theories of rationality have been based primarily upon the belief/desire psychology inherited from Hume. One of the simple consequences of the preceding account is that such a psychology is inadequate for the construction of a rational agent. A rational agent must have beliefs, situation-likings, feature-likings, intentions, primitive desires, instrumental desires, and present-tense action desires. Each of these states plays an importantly different functional role in rationality. It is particularly important to distinguish likings from desires, because most theories of rationality fail to make this distinction. The role of desires is to encode the goals whose adoption initiates acting and planning, whereas the role of likings is, on the one hand, to provide the comparisons used in selecting the goals that supply the focus for planning, and on the other hand, to provide the basis for evaluating plans.

One of the most important consequences of this account is that only an ideally situated agent, unconstrained by limited computational resources and able to engage in large amounts of epistemic reasoning without doing any practical reasoning, could rely upon ratiocination for all of its practical cognition. Any real agent must make heavy use of Q&I modules in order to produce practical conclusions within a reasonable amount of time. In other words, ratiocination is not a general solution to cognition. Rationality must be built on top of prerational mechanisms. In a fully rational agent, the output of these Q&I modules is in principle correctable by ratiocination, although it is doubtful that human beings are always capable of making full corrections.

Practical cognition begins with various kinds of evaluative attitudes. Philosophers have disagreed about whether these evaluative attitudes are
themselves subject to rational assessment. (Nagel, 1970, and Gauthier, 1986, both argued that desires can be criticized rationally; Parfit, 1984, maintained that they cannot.) It is a consequence of this theory of rationality that primitive desires, instrumental desires, feature-likings, and present-tense action desires are all subject to rational evaluation. Feature-likings are produced by a special mechanism, but their purpose is to serve as a Q&I module for finding situation types having high expected situation-likings. If it is ascertained that the expected situation-liking of a situation type is not high, then a fully rational agent will either give up the feature-liking or ignore it in practical reasoning. Obviously, instrumental desires can be criticized by criticizing the plan from which they are derived. What is perhaps more surprising is that primitive desires are also criticizable. Primitive desires can be produced either by optative dispositions or by ratiocination on the basis of the belief that a situation type has a high relative expected value. In the latter case, the desire is criticizable if the grounds for the belief are criticizable. If, instead, the desire is produced by an optative disposition, then (as we have already seen) it can be criticized if it is ascertained that the object of the desire does not have a high relative expected value (you wouldn’t like it if you had it). Some present-tense action desires are derived from plans, and these can obviously be criticized by criticizing the planning. But others are produced nonratiocinatively by optative dispositions, and it may seem that they are immune from criticism. However, recall that the objective of practical cognition is to help the agent to live a good life, where that consists of living in such a way that one likes the situation tokens one is in. The prerational optative dispositions that produce present-tense action desires can still be evaluated by asking whether they contribute to this goal. The present-tense action desires produced by these dispositions do not themselves have any automatic claim to fulfillment. It is particularly obvious that optative dispositions produced by conditioning can be harmful when judged by this standard. If it is determined that a particular desire of this sort does not contribute to leading the good life, that should serve to dispel it. If an agent retains that desire despite having this belief, the desire is irrational. (Of course, this is one respect in which human beings are not fully rational: Acquiring such beliefs does not automatically dispel the irrational desires.) In the present theory, situation-likings are immune from rational criticism. This is an automatic consequence of the conception of rationality being investigated here. By definition, rational agents seek to maximize their liking of their situations. One of the referees observed that this has consequences that may be deemed undesirable. For example, what if my liking of a situation depends very much on my being misinformed about it? However, any account that would handle this differently would, of necessity, be based upon a different conception of rationality.

Another matter on which there is no consensus is whether predicable future desires and likings should have the same status in practical reasoning
as do present desires and likings. (Nagel, 1970, argued that they should; Gauthier, 1986, and Parfit, 1984, disagreed.) For instance, I know that I am going to get hungry tomorrow, and that will lead to a desire to eat lunch, but I do not now have that desire. Should I take the predictable future desire into account in making plans for tomorrow? To me, it seems intuitively obvious that I should, but I would rather not rest my case on intuition here. In fact, it follows from my general account of practical cognition that I should take this predictable future desire into account. Let me repeat again: The objective of practical cognition is to help the agent to live a good life, where that consists of living in such a way that one likes the situation tokens one is in. The court of last appeal is situation-likings. Situation-likings pertain to liking situation tokens when one is in them. Situation-likings can only be directed at the situation token in which the agent currently finds itself. It makes no sense to ask, for example, whether the agent would like a certain situation token if the agent were in that situation but had a different conative structure. Situation tokens are fully specific. It is logically impossible to be in the same situation token but have a different conative structure. Now apply this observation to the issue of what role future likes and dislikes play in practical cognition. Come tomorrow, if I do not eat lunch, that will contribute to my disliking my situation. Accordingly, practical cognition will better serve its purpose if it takes account of this. In estimating expected situation-likings, these must be based upon what the agent thinks its conative structure is going to be like at the time in question. In evaluating a possible situation that may occur tomorrow, what is relevant is what I will like or dislike if I am in that situation, not what I like or dislike now.

As far as desires are concerned, they have no ultimate standing in practical cognition. A desire is reasonable only if it is appropriately related to expected situation-likings. So neither future desires nor present desires can affect rationality directly. However, they can do so indirectly, and insofar as a future desire is related to future likings, it has the same rational status as do present desires that are related to present likings.

Consider how this account applies to a difficult case (due to Jean Hampton, personal communication). Imagine a person who has some disease, and suppose that bouts of the disease affect him so that he desires not to be treated. On the other hand, he knows that the bouts have this effect, and so he announces to his friends that when he suffers such a bout, they should physically restrain him and force him to get treatment. We can ask two questions. First, is it rational for him to make plans now for circumventing his predictable future desire not to be treated? Second, when he is suffering a bout of the disease, is it rational for him to resist treatment? Let us make the reasonable assumption that if he is treated, he will later be glad that he was. Then it follows that living the good life requires his being treated,
because that will maximize his being in situation tokens that he likes. Accordingly, his desire not to be treated is rationally criticizable. However, this is only on the assumption that he will later be glad he was treated. Suppose instead that a single bout of the disease changes him forever so that he will not only desire not to be treated, but even if treated will continue to dislike his situation. For instance, suppose he knows that the treatment will leave him paralyzed from the neck down and in continual pain, and he knows that he will not want to live in that state. Then, on the theory constructed before, it is his present desire to circumvent his future desire that becomes irrational. This seems to me to be a congenial consequence of the theory.

A different kind of case occurs when we cannot change our situation but can change our conative structure so that we like our situation better. The present theory has the consequence that we should do that. Is this prediction correct? It seems to be that it is. For instance, consider a hostage imprisoned by terrorists. He cannot escape, but he can hypnotize himself so that he enjoys his captivity. Given that ability, it seems to me that it would obviously be a good thing for him to do.

On the theory of rationality constructed here, reasoning includes not just epistemic reasoning but also practical reasoning wherein desires are adopted on the basis of beliefs about the expected situation-likings of potential goals, intentions are adopted on the basis of beliefs about the relative values of plans, and actions are produced by choosing the strongest present-tense action desires. This contrasts with the traditional Humean picture of reasoning that recognizes only epistemic reasoning. Such a restriction seems to me to be unduly narrow. All of these state transitions among beliefs, desires, and intentions, are dictated by the functional architecture of a rational agent. They are part of the makeup of a rational agent, so why not call them "reasoning"? They are certainly part of rationality, and insofar as an agent fails to perform these state transitions, the agent is subject to the criticism that it is being irrational in the sense of not behaving as a fully rational agent would behave. We can decline to call these state transitions "reasoning" if we like, but that can be no more than a linguistic convention and does not indicate any deep difference between these state transitions and those involved in epistemic reasoning. As far as I can see, they are all on a par as being part of the functional description of a fully rational agent.

5. CONCLUSIONS

This investigation began with a very general conception of a rational agent as an agent that likes or dislikes its situation to varying degrees and when it finds the world not entirely to its liking it tries to change that. It was shown that general logical considerations and considerations of computational feasibility suffice to dictate a great deal of the structure of a rational agent.
We were led to the picture of a rational agent as a planning agent whose planning is directed by epistemic cognition and whose planning is initiated and evaluated by a complicated structure of evaluative attitudes. It is important to realize that this structure was not elicited by reflecting on human cognition, but was elicited as the only immediately apparent solution to a number of logical problems and problems of computational feasibility. Although the enterprise has been to describe rationality in general, and not human rationality, I take it that this throws considerable light on the nature of human rationality and explains why human cognition has much of the structure that it does.

REFERENCES


