Evaluative Cognition

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

1. Practical Cognition

Cognitive agents form beliefs representing the world, evaluate the world as represented, form plans for making the world more to their liking, and perform actions executing the plans. Then the cycle repeats. This is the *doxastic-conative* loop, diagrammed in figure one.[1] Both human beings and the autonomous rational agents envisaged in AI are cognitive agents in this sense. The cognition of a cognitive agent can be subdivided into two parts. *Epistemic cognition* is that kind of cognition responsible for producing and maintaining beliefs. *Practical cognition* evaluates the world, adopts plans, and initiates action. There is a massive literature both in philosophy and artificial intelligence concerning various aspects of epistemic cognition, and large parts of it are well understood. Practical cognition is less well understood. We can usefully divide practical cognition into five parts: (1) the evaluation of the world as represented by the agent's beliefs, (2) the adoption of goals for changing it, (3) the construction of plans for achieving goals, (4) the adoption of plans, and (5) the execution of plans. There is a substantial literature in AI concerning the construction and execution of plans, and I will say nothing further about those topics here. This paper will focus on the evaluative aspects of practical cognition. Evaluation plays an essential role in both goal selection and plan adoption. My concern here is the investigation of evaluation as a cognitive enterprise performed by cognitive agents. I am interested both in how it is performed in human beings and how it might be performed in artificial rational agents.

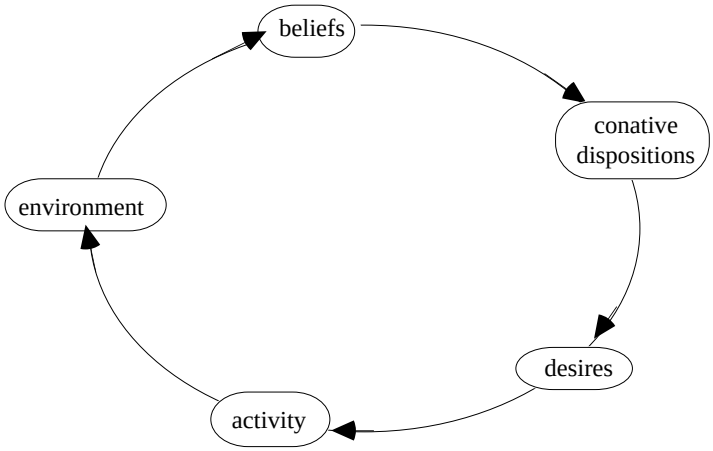

Figure 1. The Doxastic-Conative Loop

This work was supported by NSF grants nos. IRI-9634106 and IIS-0080888.

[1] The doxastic-conative loop was introduced in my (1993), and forms the basis for the theory of rational agency in Pollock (1995).

This enterprise seems on the surface to be connected with the philosophical investigation of value. In one sense, what evaluative cognition produces is assessments of value. However, this may or may not be the same concept of value that is the focus of interest in value theory. Value-theoretic investigations are predominantly metaphysical. This paper pursues the epistemology and cognitive science of value rather than its metaphysics. More specifically, this paper is concerned with those aspects of rational cognition that are concerned with the comparative assessment of competing plans and goals, and for present purposes *value* is simply defined to be whatever such assessments measure. Thus rather than starting from a metaphysics of value and asking how we can learn about values, I start with an investigation of certain aspects of cognition. The metaphysically inclined can go on to ask about the nature of the values that are the objects of such cognition and how they related to other philosophical enterprises, but this paper will not pursue those matters.

There is an immense philosophical literature on values, but for the reasons just given, most of that work is not easily applicable to the investigations of the cognitive scientist. There is a small but growing literature on decision-theoretic planning in AI that makes use of values in directing agents' activities, but it tends to be mute on the source of the values. In AI it is generally just assumed that the agent can attach values to states of affairs and then those values are used in directing plan adoption and execution. My interest here is in the value assessments themselves. Where do cognitive agents get the evaluations they employ in practical cognition? How are values computed? It turns out that purely computational considerations can take us a long way towards answering these questions.

1.1 The role of values

It is useful to begin by sketching how values are used in practical cognition. This will constrain theories of evaluative cognition. Values play two essentially different roles. First, they are used in selecting goals. Goals are *ways we want the world to be*. More technically, they are *situation-types*. *Situation tokens* are "total" ways the world might be, i.e., complete specifications of possible worlds. Situation-types abstract from the situation tokens and are partial descriptions of ways the world might be. Practical cognition aims at changing the world so that it exemplifies a situation-type the agent values. Practical cognition begins by selecting goals, which are valued situation-types the agent believes the world either will not or may not come to exemplify unless the agent intervenes in some way. So values are used in selecting goals.

The second use of values in practical cognition is in deciding what plans to adopt. Plans are constructed that will, with appropriate probabilities, achieve goals. But even if a plan can be expected to achieve its goal, it is not automatically a good plan. This is because plans do more than achieve their goals. They have side effects. Some of the side effects are of negative value, and can be labeled *execution costs*. There can also be fortuitous side effects, wherein the plan produces effects of positive value without having originally been designed with those as goals.

I assume that the evaluation of plans is in some sense decision-theoretic, but the details of that evaluation are complex. This is in part because the context in which a plan is executed can affect its expected-value, both by affecting the probabilities of outcomes and by affecting the values of the outcomes. Among other things, the context in which a plan is executed will be affected by what other plans the agent adopts and executes. Accordingly, a plan cannot be evaluated in isolation from the agent's other plans. The agent's entire set of plans must be evaluated as one package, and the decision whether to adopt a new plan must be based upon its effect on the agent's entire set of plans. I will say more about this in section six.

1.2 Goal-directed planning

The two ways in which values are used in practical cognition are strangely disparate. The ultimate objective of practical cognition is to adopt and execute plans, so it seems that the

primary function of evaluation should be the evaluation of plans. Plans are evaluated in terms of all of their possible effects, not just their goals. Furthermore, the value of a goal in the context of a plan may be different from its value in isolation. For example, an agent might begin with the goal of having a dish of vanilla ice cream. A playful friend might offer to provide the ice cream if the agent will eat a dill pickle first. That produces a plan for achieving the goal, but it is not a very good plan because the value of eating the ice cream is significantly diminished in the context in which it is preceded by eating a dill pickle.

If the value of a plan is not a function of the value of its goals (in isolation), what good are goals? Why not just construct plans directly, evaluating them decision-theoretically? There is an approach to planning that tries to do just that. MDP's (Markov decision planners), and more generally POMDP's (partially observable Markov decision planners) proceed by building a *decision tree*.[2] This is a graph in which nodes represent possible states of the world and links between nodes correspond to actions that would move the world from one state to another with some specified probability. It is assumed that we have a valuation function assigning a value to each node, and then the objective is to construct an *optimal policy*, which is in effect a global plan prescribing the best action to perform in each possible state. MDP's and POMDP's proceed by building the entire decision tree and then searching for an optimal path through it. As such, they have no use for goals. The only evaluations are of the total states of the world as represented by the nodes of the tree.

A generally recognized problem for Markov decision planning is that it is computationally infeasible in any but the simplest environments. States of the real world are characterized by a huge number of variables. To estimate the complexity of the real world, it has been estimated that there are 10^{78} elementary particles. If we take the state of a particle to be determined by four quantum states each having two possible values (a gross underestimation), each particle can be in 16 states, and so there are $16^{10^{78}}$ states of the universe.[3] This is a bigger number than we can write in non-exponential form. It would be longer than the number of elementary particles in the universe. Clearly, a cognitive agent cannot be expected to find an optimal policy prescribing actions for all of these different states. A computationally feasible policy must abstract from the true complexity of the universe, making the assumption that most differences between states do not make any difference to how the agent should behave. Suppose we could confine our attention to just 300 two-valued variables. That is pretty unrealistic—it seems clear that many more than 300 parameters can make a difference to optimal behavior, and many of them are continuous-valued rather than two-valued. But even if we could confine our attention to 300 two-valued variables, an optimal policy would have to distinguish between 2^{300} states and prescribe behavior for each. 2^{300} is approximately equal to 10^{90}, which is twelve orders of magnitude larger than the number of elementary particles in the universe. Clearly, a real agent cannot deal with policies that large, and even such policies would be inadequate because in some cases they would fail to make crucial distinctions.

Because it is generally impossible for a cognitive agent to construct (or even represent) optimal policies, it must instead abstract from the total state of the universe, identifying manipulable constituents of that state whose change tends to alter the value of the total state. The agent can then affect the value of the total state by affecting these constituents. That becomes the immediate target of practical cognition. Affecting the constituents in various ways becomes a goal, and the agent constructs plans for achieving such goals. Viewed in this light, the importance of goals lies in the role they play as part of a control structure for practical cognition. We cannot make

---

[2] A good survey of this approach can be found in Boutilier, Dean, and Hanks (1999).

[3] In fact, some of the parameters, like position, are continuous-valued, so there are really infinitely many possible states.

progress in practical deliberation by generating plans at random and evaluating them decision-theoretically, because there are too many candidates. That is tantamount to considering arbitrary partial paths through a decision tree, and is no more computationally feasible than Markov decision planning. By focusing on goals and employing planning procedures that produce plans for achieving those goals,[4] the agent is able to produce plans that have some presumption in favor of having positive expected-values. If a plan can be expected to achieve its goal, the plan can be expected (defeasibly) to have a positive expected-value unless execution costs overwhelm the value of the goal.

1.3 Cardinal measures of value

The two uses of values, in selecting goals and adopting plans, impose different requirements on values. Goals can be selected simply on the basis of their having positive value. This requires at most an ordinal measure of value. But evaluating plans decision-theoretically involves multiplying values and probabilities and summing the results. For that to make sense, there must be a cardinal measure of value. Furthermore, it is the agent that is deciding what plans to adopt, so it is not enough for the cardinal measure of value to simply exist—the agent must have cognitive access to it in order to perform decision-theoretic evaluations of plans. In other words, the agent must be able to compute values in a way that makes decision-theoretic evaluations possible. Let us turn then to the question how that can be done in general (in any cognitive agent), and more specifically how it is done in human beings.

2. Preference Rankings

We might begin by considering an answer suggested by rational choice theory, which underlies much of modern economics. That theory begins with the observation that although decision-theory requires a cardinal measure of value, human beings are unable to assign numbers to values simply by introspecting. On the other hand, humans can introspect *preferences*. That is, they can tell introspectively that they prefer one situation-type to another. Assuming that our preferences are transitive, this generates a preference ranking of all the items between which we have preferences. Frank Ramsey (1926) and Leonard Savage (1956) showed independently that if our preference ranking includes situation-types consisting of our being offered certain wagers, and the ranking satisfies some plausible axioms, then it is possible to generate a unique cardinal measure[5] which supports decision-theoretic reasoning.

In rational choice theory, the point of this technical result is to establish the existence of a cardinal measure. It must be emphasized that it shows nothing directly about how human beings perform practical cognition. In particular, it clearly does not show that humans proceed by recovering this cardinal measure from their preference rankings and then use it to reason decision-theoretically. Even if this cardinal measure exists, humans rarely assign numbers to their values and accordingly they rarely engage in explicit decision-theoretic reasoning.

Let us ask a different question. Regardless of how humans work, *could there be* rational agents that worked in this manner? The proposal would be that the fundamental value-theoretic data structure to which they appeal is a preference ranking, and a cardinal measure is computed on the basis of that preference ranking and used for subsequent decision-theoretic reasoning. I will now argue that this is impossible. An agent could not be built that works this way in a

---

[4] This is known as *goal-regression planning*, and has a long history in AI. The logic of the kind of plan-search involved in goal-regression planning is investigated in my (1998).

[5] Unique up to linear transformation.

complex environment. To see this, consider how many situation-types would have to be included in the preference ranking. Perhaps every situation-type should be included. But there are at least as many situation-types as there are states of the world that would have to be included in a decision-tree for Markov decision planning.[6] Thus the preference ranking would have to rank more situation-types than there are elementary particles in the universe. Clearly, such a preference ranking cannot constitute a primitive data structure in a real agent.

Perhaps we don't have to include all situation-types in a preference ranking. It seems reasonable to propose that the preference ranking need only include situation-types that have nonzero value. An agent is apt to be indifferent to most situation-types, so this seems to produce a markedly smaller preference ranking. But it is still not small enough. Let *P* and *Q* be situation-types, where *P* is "value-laden", i.e., has a nonzero value, and *Q* is not. Then *P* will be included in the preference ranking and *Q* will not. However, unless *Q* interacts with *P* in such a way as to cancel its value, (*P*&*Q*) will also be value-laden, and so if all value-laden situation-types are included in the preference ranking, (*P*&*Q*) must be included. Hence little is gained by leaving *Q* out of the preference ranking. To illustrate, suppose again (unrealistically) that states of the world can be characterized by just 300 two-valued parameters. Then there will be 600 "simple" situation-types each consisting of one of these parameters having a specific value. All other situation-types will correspond to conjunctions of these simple situation-types. Suppose (again unrealistically) that just 30 of these simple situation-types are value-laden, and suppose that compound situation-types are value-laden only by virtue of containing one or more value-laden simple situation-types as constituents. There will be 2^{600} complex situation-types, but they need not all be ranked. However, the only situation-types that need not be ranked are those containing no value-laden constituents. There will be 2^{540} of these. $2^{540}/2^{600} = 2^{-60} = .000000000000000008$. Thus only a miniscule proportion of the situation-types are omitted from the ranking.

Perhaps we can simplify the preference ranking still further. Presumably it will usually be true that if *P* is value-laden and *Q* is not, then the value of (*P*&*Q*) will be the same as that of *P*. In that case we can simply leave (*P*&*Q*) out of the ranking, and when the time comes to compare it with other situation-types we *compute* a preference by taking it to be the same as the preference for *P*. We cannot always omit conjunctions (*P*&*Q*) from the preference ranking, because sometimes *P* and *Q* will interact in such a way that the value of (*P*&*Q*) is different from that of *P*, but we can record that fact by including (*P*&*Q*) in the preference ranking. Using this strategy, when a conjunction is not contained in the preference ranking and we want to compute a preference between it and some other situation-type, we can do that by identifying its place in the ranking with that of the largest conjunction of a subset of its conjuncts that is explicitly contained in the ranking. There may be more than one such conjunction, but they will all be ranked alike if the absence of the conjunction from the ranking means that it has the same value as any smaller conjunction from which it can be obtained by adding value-neutral conjuncts.

This strategy achieves a significant decrease in the size of the preference ranking. In the above example we would only have to include 2^{60} elements in the ranking. However, that is still a pretty large number, approximately 8×10^{19}. The best current estimates are that that is many orders of magnitude larger than the entire storage capacity of the human brain,[7] and that is from just 30 value-laden simple situation-types. Realistically, human beings must be faced with at least 150 value-laden simple situation-types (and probably orders of magnitude more). That produces a preference ranking containing at least 2^{300} items, which is again larger than the number of particles in the universe. This cannot possibly be the way human beings record values.

---

[6] In fact, if there are *N* states of the world then there are at least 2^N situation-types.

[7] See Landauer (1986) in which it is argued that a number of different techniques converge on a figure of around 10^9 bits.

I think it must be concluded that the fundamental value-theoretic data structure in human beings does not take the form of a preference-ranking. This is not to say that humans don't have the requisite preferences, but just that they are computed from something else that can be stored more compactly. What might that be? There is a simple answer—just store numerical assignments of value to value-laden situation-types. Preferences can then be computed by comparing the numbers. But is this really more compact? If we have to store a numerical value for every value-laden situation-type, we are no better off than with preference rankings. We don't really have to store numbers for *every* situation-type. Just as in the case of preference-rankings, we can take the absence of a situation-type from the database to signify that its value is the same as that of the largest conjunction of its conjuncts that is present in the database. But still, that makes the database no smaller than the smallest preference ranking we were able to produce above.

However, for a database of numbers, if the numbers represent a cardinal measure, a considerable additional simplification can be achieved. Just as we can assume defeasibly that conjoining a value-laden situation-type with a value-neutral one will produce a conjunction whose value is the same as that of the value-laden conjunct, so it seems reasonable to assume defeasibly that conjoining two value-laden "simple" situation-types will produce a conjunction whose value is the sum of the values of its conjuncts. In that case, it can be omitted from the database and a value computed for it as necessary. Of course, value-laden situation-types are not always independent of one another. Recall the earlier example of eating vanilla ice cream after eating a dill pickle. But in that case we can record the lack of independence by including an assignment to the conjunction in the database. Let us say that P and Q are *value-theoretically independent* iff the value of (P&Q) is the sum of the value of P and the value of Q. On the assumption that simple situation-types are usually value-theoretically independent, a database recording values produced by 150 value-laden simple situation-types will require on the order of 300 entries, as opposed to the 2^{300} entries required in a preference ranking. This difference is the difference between the trivial and the impossible.

It is worth noting why the same simplification cannot be achieved using preference rankings. It might be supposed that if we can assume defeasibly that value-laden states are value-theoretically independent then we need not store most conjunctions in the preference ranking. Can't we compute the position of the conjunction from the position of the conjuncts? The answer is that we cannot. The preference ranking gives us only an ordinal measure of value. To compute the value of a conjunction from the values of its conjuncts, we have to be able to add values, which requires a cardinal measure.

The preceding discussion was predicated on the assumption that the simple value-laden situation-types are produced by setting the values of two-valued parameters. At least some of the relevant parameters, like position, are continuous-valued. Position is probably not itself a value-laden situation-type, but when combined with other value-laden situation-types it can change their values. An ice cream cone on the moon is not nearly so valuable as one in my hand. Continuous-valued parameters make the preference ranking infinite, in which case it clearly cannot constitute the evaluative data structure underlying the value computations of either human beings or artificial rational agents. On the other hand, continuous-valued parameters need create no difficulty for storing numerical values. Rather than storing a constant value, we simply store a function of the parameter.

The upshot is that storing a cardinal measure of values for simple situation-types is a vastly more efficient way of storing values than storing them in the form of a preference ranking. The cardinal measure allows us to use arithmetic in computing the values of composite situation-types, and that in turn allows us to omit the computable values from the value-theoretic database. I will turn to the details of this computation in the next section.

3. Storing Values

The question addressed in this section is how to efficiently store values. The proposal is to store only those values that cannot be computed on the basis of other stored values. I assume that the situation-types that are to be evaluated can contain one another as constituents and I represent that with conjunction. If P contains Q as a constituent, then P is a conjunction of situation-types and Q is one of its conjuncts. This imposes rather narrow constraints on the logical form of the situation-types to be evaluated. They cannot, for example, be disjunctions of one another. This restriction does not seem to me to be inappropriate. We *can* evaluate disjunctions, but only in terms of their expected-values. That is, the evaluation of $(P \vee Q)$ would be its expected-value:

$$V(P)\cdot\text{PROB}(P / P\vee Q) + V(Q\&\sim P)\cdot\text{PROB}(Q\&\sim P / P\vee Q).$$

The situation-types to which values are attached directly will be those in terms of which expected-values are computed, and these can always be regarded as conjunctive descriptions of the world. Let us call them *state descriptions*. Situation-types that are not state descriptions must be evaluated in terms of expected-values rather than directly.

Value computations for state descriptions will be based upon numerical values stored in an *evaluative database* of "primitive values". I will return below to the question where the primitive values come from. The states assigned values by the evaluative database will be called "primitively value-laden states". Because values can vary with context, some of these primitively value-laden states will be conjunctions having others as conjuncts. E.g., there may be one value assigned to *eating a bowl of vanilla ice cream*, and another value assigned to *eating a bowl of vanilla ice cream shortly after eating a dill pickle*. The primitively value-laden states will also be allowed to contain free variables, corresponding to parameters that affect the value, and the value associated will be a function of those free variables rather than a constant.

The simple non-conjunctive state descriptions out of which others are constructed by conjunction will be called "simple states". This is a value-theoretic concept, not an ontological one. That a state is simple is not a comment on the metaphysical structure of the world, but just a comment about the structure of our evaluative database. That a state is simple tells us only that our evaluative database does not assign values to any of its logically simpler constituents.

Two states P and Q are value-theoretically independent iff $V(P\&Q) = V(P)+V(Q)$. The intent is that the evaluative database will contain a conjunction of states only when the value of the conjunction cannot be computed as the sum of the values of some of its conjuncts. To make this precise, we must state the rules for computing values for state descriptions that are not assigned values directly.

I will take conjunctions to be of arbitrary length, and the conjuncts of a conjunction $(P_1\&...\&P_n)$ will be $P_1,...,P_n$. A *subconjunction* of $(P_1\&...\&P_n)$ will be any conjunct or conjunction of conjuncts of $(P_1\&...\&P_n)$. Note that $(P_1\&...\&P_n)$ is one of its own subconjunctions. I will also identify the conjunction of two conjunctions with the conjunction of their conjuncts.

Turning to the rules for computing values for state descriptions, the simplest case is a state description S that has no subconjunction which is assigned a value directly. This signifies that the state description is value-neutral, i.e., $V(S) = 0$.

Consider a state description $(P_1\&P_2)$ where P_1 and P_2 are simple states. If P_1 is assigned a value $V(P_1)$ directly, and P_2 is not assigned a value, then $V(P_2) = 0$ and the presumption is that P_1 and P_2 are value-theoretically independent, in which case $V(P_1\&P_2) = V(P_1)+V(P_2) = V(P_1)$. This default computation will only be overridden if $(P_1\&P_2)$ is assigned a value directly.

Consider a state-description $(P_1\&...\&P_n)$ where $P_1,...,P_n$ are simple states. If just one subconjunction of $(P_1\&...\&P_n)$ is assigned a value, then as above that will also be the value

computed for (P_1&...&P_n). But suppose instead that several subconjunctions are assigned values. We can distinguish several cases:

(a) There might be one primitively value-laden subconjunction *S subsuming* all other primitively value-laden subconjunctions of (P_1&...&P_n), in the sense that the set of conjuncts of any other such subconjunction will be a subset of the set of conjuncts of *S*. Precisely:

DEFINITION: (A_1&...&A_n) *subsumes* (B_1&...&B_m) iff $\{B_1,...,B_m\} \subseteq \{A_1,...,A_n\}$.

If *S* is a primitively value-laden subconjunction of (P_1&...&P_n) that subsumes all other primitively value-laden subconjunctions, then the value assigned to *S* will override the values assigned to any other subconjunctions, and so V(P_1&...&P_n) = V(*S*).

(b) There might be several primitively value-laden subconjunctions $S_1,...,S_k$ such that (i) every primitively value-laden subconjunction of (P_1&...&P_n) is subsumed by some S_i, and (ii) the S_i's have no conjuncts in common. This should signify that the S_i's are value-theoretically independent and hence we should have V(P_1&...&P_n) = V(S_1)+...+V(S_k).

(c) Suppose (P_1&...&P_n) has two primitively value-laden subconjunctions S_1 and S_2 neither of which subsumes the other, but suppose S_1 and S_2 have the conjunct *P* in common. To identify V(P_1&...&P_n) with V(S_1)+V(S_2) is to double count any contribution from *P*. If V(*P*) = 0, this is not a problem, but what happens when V(*P*) ≠ 0? Notice that this is not a substantive question about value. Rather, it is a question about how best to organize the evaluative database in the interest of compactness. We could simply rule that in this case (P_1&...&P_n) must be assigned a value directly, but additional compactness can be achieved by making some default assumptions. The value of a state *P* in a context *C* may be different from the value of *P* simpliciter. Let us write this as V(*P* / *C*). We can identify this with the value *P* contributes to the conjunction (*P*&*C*) over and above the value of *C*, and define it precisely as follows:

DEFINITION: V(*P* / *C*) = V(*P*&*C*) – V(*C*).

Then we can extend the presumption of value-theoretic independence by assuming that V(P_1&...&P_n) = V(S_1 / *P*) + V(S_2 / *P*) + V(*P*) = V(S_1) + V(S_2) – V(*P*).

(d) More generally, suppose (P_1&...&P_n) has several primitively value-laden subconjunctions $S_1,...,S_k$ none of which subsumes another, but some of which have conjuncts in common. For example, suppose *A*&*B*&*D*, *A*&*C*&*D*, *B*&*C*&*D*, *A*&*D*, *B*&*D*, *C*&*D*, and *D* are primitively value-laden, and consider V(*A*&*B*&*C*&*D*). This should be

$$V(A\&B\&D) + V(A\&C\&D) + V(B\&C\&D) - V(A\&D) - V(B\&D) - V(C\&D) + V(D).$$

Note that the final term is required to avoid triple-counting the contribution of *D* in the subtraction. Let $S_i \cap S_j$ be the conjunction of the conjuncts S_i and S_j have in common. Then in general we should have:

$$V(P_1\&...\&P_n) = V(S_1)+...+V(S_k) - \sum_{i\neq j} V(S_i \cap S_j) + \sum_{i\neq j, i\neq k, j\neq k} V(S_i \cap S_j \cap S_k) - ...$$

where we subtract the values of pairs, add the values of triples, subtract the values of quadruples, etc. A more compact way of expressing this will be presented in section seven, where it will be shown that the database calculation can be derived from general features of cardinal measures.

(e) Finally, in the most general case, (P_1&...&P_n) might be as in (d) except that some of the primitively value-laden subconjunctions are subsumed by others. In that case the subsuming subconjunctions take precedence. So let $S_1,...,S_k$ be the primitively value-laden subconjunctions of (P_1&...&P_n) that are not subsumed by other primitively value-laden subconjunctions. Then:

$$V(P_1\&...\&P_n) = V(S_1)+...+V(S_k) - \sum_{i\neq j} V(S_i \cap S_j) + \sum_{i\neq j, i\neq k, j\neq k} V(S_i \cap S_j \cap S_k) - ...$$

I will refer to the preceding calculation as the *database calculation*. It is to be emphasized that this is a principle for retrieving values from the evaluative database. It is not a substantive principle about values, but rather an organizational principle for the evaluative database. If the actual value of $(P_1\&...\&P_n)$ does not accord with this computation, then it must be included as a primitive entry in the database. On the other hand, there is a substantive assumption underlying this organization of the evaluative database. Organizing it in this way will only achieve compactness if the values of state descriptions are normally related in accordance with the database calculation. The justification of this assumption will be addressed in section seven.

4. The Source of the Primitive Values

Goals and plans can be evaluated decision-theoretically by using probabilities and the values of state descriptions to compute expected-values for situation-types that are logically more complex than state descriptions. The values of state descriptions are, in turn, computed as above on the basis of the evaluative database. The entries in the evaluative database represent *primitive values*, in the sense that they cannot be derived (via the database calculation) from other values. Where do the primitive values come from?

By definition, primitive values cannot be derived from other values in the evaluative database. That leaves two possibilities. Either primitive values are primitive constituents of an agent's cognitive architecture, or there is some cognitive mechanism enabling the agent to infer (derive) the values from something else. If the values are to be derived from something else, the obvious suggestion is that they be derived from facts about the world that the agent is able to learn by epistemic cognition. At this point, some philosophers may wave their hands and proclaim that primitive values represent "objective goods", where this is some sort of metaphysical notion. I do not understand this claim well enough to deny it, but regardless of whether it is true it is not much help to the cognitive scientist trying to understand the role of value in cognition. An appeal to objective goods can only contribute to an understanding of evaluative cognition if it is accompanied by an explanation of how an agent can learn about objective goods. No plausible answer to this question has ever been proposed.[8]

As uninformative as an appeal to objective goods is, it is the only kind of answer that comes to mind as an attempt to ground evaluation on epistemic conclusions. I presume then that values cannot be grounded on epistemic conclusions. This suggests that the evaluative database is a primitive constituent of the agent's cognitive architecture. The elements of the database are not derived from anything else. If our interest is in building artificial rational agents, it seems we could stop here. This account tells us how evaluative cognition in such agents can work. We just assign some primitive values and give the agent the ability to compute other values in the manner described above.

However, there is a deep problem for this way of grounding evaluative cognition. The elements of the evaluative database are assignments of value to situation-types, and as the agent uses these assignments for further reasoning about the expected-values of goals and plans the

[8] I am inclined to think that this approach puts the cart before the horse. If objective goods make sense, they are best explained by first giving an account of value-theoretic cognition and then explaining objective goods as those things that are revealed by that cognition to have value. Little progress can be made by taking objective goods as basic and then trying to understand value-theoretic cognition as whatever is required to find out what is objectively good.

agent must have cognitive access to the elements of the database. In particular, the agent must have a system of internal (mental) representations enabling it to think about the situation-types that are assigned values. For a rather simple cognitive agent operating in a narrowly circumscribed environment, that may not be a problem, but for a truly sophisticated cognitive agent capable of functioning in a wide variety of environments, this is a major difficulty. To take an extreme example, suppose we want the agent to value democracy. To incorporate this into the agent's evaluative database, we must equip the agent with the concept of democracy. It is far from clear what that involves, but it seems extraordinarily unlikely that the agent could have that concept without having substantial knowledge of the world. Of course, "democracy" is an extreme case, but it seems unlikely that more mundane concepts like "mother" or "father" will prove much simpler. For artificial agents, we might provide the requisite knowledge of the world by building an elaborate "a priori" world model into the agent. But there are two problems with that approach. First, building the world model is a formidable task—probably impractically so.[9] Second, the resulting cognitive agent will be brittle, in the sense that it will be unable to function in a world differing at all from its built-in world model. Clearly, human beings do not work in this way. They acquire their world model through learning, and they acquire the concepts required to think about the world as part of that learning. I cannot prove this without a general theory of concepts, but it seems likely that any general-purpose cognitive agent capable of functioning in unconstrained environments must work similarly. This implies that either the concepts employed in the agent's evaluative database are extremely simple ones that do not require extensive world knowledge or the evaluative database is acquired along with the agent's world knowledge rather than being built in from the beginning. I assume that restricting the evaluative database to simple concepts will not be satisfactory in a sophisticated agent, so let us explore the other alternative. The suggestion is that the evaluative database is constructed incrementally as the agent acquires knowledge of the world. The incremental construction cannot be random—it must be based upon the agent's discoveries about the way the world works. In other words, there must be a variety of rational cognition whose purpose is to add elements to the evaluative database in response to acquiring knowledge of the world. This rational cognition can be viewed as a kind of inference, in which case its operation has the function of deriving the values encoded in the evaluative database from something else.[10]

It was argued above that the primitive values encoded in the evaluative database cannot be derived from epistemic conclusions, but I have now argued that they must be derived from something. If they cannot be derived from epistemic conclusions, what is left? I will suggest an answer to this question by looking more closely at human evaluative cognition.

5. Grounding Values in Human Cognition

The primitive elements of the evaluative database must be derived from something, but they cannot be derived from epistemic conclusions. My proposal will be that in human beings they are derived from other values, but the values in question are of a logically different kind from the values encoded in the evaluative database. Primitive values were defined to be the values stored as elements of the evaluative database, but I am now suggesting that there are more basic values from which the primitive values themselves are derived. *Basic values* are values that cannot be derived from any values but from which other values can be derived. So

---

[9] Consider, for example, the difficulties encountered by the Cyc project (http://www.cyc.com), which is trying to build just such a world model.

[10] This is what Millgram (1997) calls "practical induction".

my suggestion is that in human beings there are values more basic than the primitive values. To defend this proposal, let us begin by asking, "What are the basic values in terms of which human evaluative cognition proceeds, and how are they encoded in human cognition?" As we will see, this is complicated by the fact that humans exhibit several different conative processes, and it is not initially evident how they fit into the theoretical picture of evaluative cognition adumbrated above.

5.1 Desires

Let us begin with desires. Human beings form desires for various situation-types and then try to achieve them. That is just to say that the desires encode goals which the agent then tries to achieve. Do desires also encode basic values in human beings? David Hume thought so, but I doubt it, for two different reasons. The simplest reason is that desires can be based on false expectations regarding the situation-type desired and we regard that as a reason for changing the desire. For example, a woman might think that she would really enjoy foreign travel, and on that basis form the strong desire to engage in it. But when she achieves her goal, she might discover that it is not at all the way she imagined—she is bored by the tedious airplane trips, she hates the unfamiliar food, and is frightened by being surrounded by strangers speaking a foreign language. The desire for foreign travel quickly evaporates when she discovers what foreign travel is really like, and we would regard her as irrational if it did not. This strongly suggests that desires do not encode basic values, i.e., they do not represent the starting point for evaluative cognition. In retrospect, this seems hardly surprising. Surely foreign travel is not the sort of thing that will have a value assigned to it directly. Rather, foreign travel is a logically complex situation-type that ought to be evaluated by computing an expected-value. That is, it has different possible outcomes, each having some probability of occurrence, and foreign travel should be evaluated by evaluating those outcomes and discounting them by their probabilities. Furthermore, it is a factual question what those possible outcomes and probabilities are, and one can easily be wrong about that. This is at least part of the mistake being made by the woman who initially desires foreign travel but later discovers that she dislikes it.

It is useful to distinguish between *intrinsic values*—things that are valued in and of themselves—and *instrumental values*—things that are valued because they have a tendency to bring about other things having intrinsic value. Foreign travel is not the sort of thing that we would normally expect to have intrinsic value. Rather, the value of foreign travel resides in its expected tendency to bring about other simpler situation-types that we think we would like in and of themselves. These might include seeing beautiful sights, meeting interesting people, tasting novel food, and so forth. It seems clear that most of our desires are for things that have only instrumental value. Basic values, on the other hand, ascribe intrinsic value to their objects. So human desires do not automatically encode basic values.

We can make a distinction within desires that parallels the distinction between instrumental and intrinsic values. Some desires are acquired as a result of means-end reasoning, and the objects of these desires are desired only instrumentally, as a way of achieving other non-instrumental desires. It might be proposed that although desires in general do not encode basic desires, non-instrumental desires do.

However, there is a way in which even non-instrumental desires can be rationally criticized as getting things wrong. One of the reasons our foreign traveler thought she would enjoy traveling is that she expected to like eating foreign food. In fact, she hated it. She was making a factual mistake here, but it was not the same kind of factual mistake discussed in the first objection. When she failed to predict the boring aspects of foreign travel it was because in imagining what foreign travel would be like she simply overlooked some of the probable outcomes, e.g., long plane flights in cramped seats followed by several days of jet lag. But in the case of eating foreign food, she wasn't overlooking anything. She was just wrong about whether she

would like it. This highlights a human conative process that is distinct from desire—*liking* being in a situation-type. The distinction is an important one, because it seems that desires are rationally criticizable by appeal to likings. If we desire something but know that we wouldn't like it if we got it, we regard that as a criticism of the desire.

In ethics, most desire theories follow Sidgewick in being *informed* desire theories. According to such theories, value attaches to what one *would* desire if one were fully informed about all relevant matters. It is worth noting that the preceding problem can arise even for fully informed desires. Compulsions are normally desires one should not have. Knowing that one should not have a certain compulsion is not sufficient to make it go away. Thus what one *should* (rationally) desire need not be the same as what one would desire if one were fully informed. In particular, one can irrationally continue to desire something even when knowing one would not like it if one got it.

5.2 Feature-Likings

The kinds of situation-types that we like or dislike are typically characterized by some features of the situation-tokens exemplifying that type. For example, I like eating Greek food. I will refer to this conative attitude as *feature-liking.* Technically, feature-liking is a propositional attitude. That is, it is an intensional state whose object is the situation-type liked or disliked. The intensional state includes a representation of the situation-type that is its object.

Desires are a bit like predictions of feature-likings. Typically, we desire something and adopt it as a goal because we expect to like it when we get it. However, human desires cannot literally be identified with predictions (i.e., beliefs) abut feature-likings. As just noted, sometimes we persist in having desires for things we know we will not like if we get them. Some compulsions are like this. One might find oneself drawn irresistibly to a member of the opposite sex even while knowing that if the desire is achieved, they will live to regret it. Or more simply, the smell of a garlic laden dish may create an overwhelming desire to partake of it even though one knows that it won't taste as good as they imagine and the ensuing gastrointestinal distress will make them regret their gustatory indiscretion. Conversely, one can know that one would like something if it happened, but fail to desire it. For example, I know that I always enjoy skiing when I go, but for some reason that knowledge does not generate a desire to go skiing. It ought to, and its failure to do so is a mark of irrationality.[11] So, desires are not the same thing as predictions of feature-likings. However, functionally, desires play much the same role as predictions of feature-likings and they are rationally criticizable by the same considerations that would lead us to regard such predictions as faulty. It is their predictive flavor that makes desires inappropriate candidates for the source of basic (intrinsic) values.

What about feature-likings? Might they be the source of basic values? The first of the two objections to the proposal that desires represent the starting point for evaluative cognition applies equally to feature-likings. There is a sense in which feature-likings are also predictive and can be mistaken. Consider our foreign traveler again. She begins with a strong desire to engage in foreign travel, and eventually she finds herself in a financial position to achieve her goal. For some years she becomes a globe trotter, traveling to numerous foreign destinations. In fact, she never enjoys her travels, but it may take some time for her to realize that, and in the meantime if you ask her, "Do you like foreign travel" she will reply, "Oh yes, I really like it". What are we to make of this answer? I think there is a sense in which she is right—she does like foreign travel—and another sense in which she is wrong. The sense in which she is right is that she is in

---

[11] Tom Christiano has pointed out to me that we sometimes employ secondary mechanisms to get ourselves to desire doing things that we think we should do but don't naturally desire doing. For example, if a friend suggests we go skiing together, I may concur knowing that my desire not to disappoint my friend will get me onto the slopes and then I will enjoy myself.

the intensional state of liking that situation type. This is just a remark about her current psychological state. The sense in which she is wrong is that she does not have a disposition to enjoy herself when she engages in foreign travel. Although the latter is true, she does not notice that it is and so retains the feature-liking. To say that she enjoys (or does not enjoy) foreign travel in the dispositional sense is to assert a statistical generalization about her. It is to say that engaging in foreign travel *tends to cause her* to enjoy herself. One can be ignorant of this sort of causal generalization about oneself.

Much like desire, the intensional state of feature-liking is functionally similar to *believing* that the situation-type that is the object of the liking will tend to be conducive to liking your current situation, and it is rationally criticizable by appeal to the same considerations that would make the belief rationally criticizable. But also like desire, feature-liking is a different psychological state from the belief.

If you like a situation-type, in the sense of having a feature-liking for it, but you know that being in situations of that type invariably makes you unhappy, then should you regard being in situations of that type as having positive value? It doesn't seem so. It seems that you should regard it as having negative value and regard your feature-liking is in some sense "in error". This seems to indicate that feature-likings are not the starting point for evaluative cognition either. One can make the same point about feature-likings that was made above about desires, viz., feature-likings are, in effect, decision-theoretic evaluations. They are concerned with instrumental values rather than intrinsic values.

5.3 Situation-Likings

When our foreign traveler engages in foreign travel but fails to enjoy it, she dislikes her current situation. In human beings, such *situation-liking* seems to provide the court of last appeal in evaluative cognition. If we either like or desire a situation-type, but being in situation-tokens of that type does not contribute causally to our liking them, then it seems that our likes and desires are rationally criticizable, and we should not regard achieving them as contributing value to our situation. A rational agent who realizes that these likes and desires are not conducive to situation-liking should not pursue them.

What kind of a state is situation-liking? I described it as "liking one's current situation". That suggests that it is an intensional state. A difficulty arises, however, when we try to say what the object of this intensional state is. What counts as "one's current situation"? One natural suggestion is that it is the possible world in which one resides. The main problem for this view is that the possible world does not change over time. It includes all truths, past present and future. But your situation-liking can change over time. Sometimes this is because your beliefs change, so you believe yourself to reside in a different possible world than you previously believed yourself to be resident in. However, your situation-liking can change *even without changes in your belief*. If I am currently in the dental chair having a cavity filled, I may like my situation less than I will like it tomorrow, despite the fact that tomorrow I will remember my encounter with the dentist and still regard myself as resident of the same possible world.

We might try to accommodate the temporal variability of situation-liking by proposing that the object of situation-liking is not a possible world after all, but a *time slice* of a possible world, where that consists of just what is true in the world at a particular instant. The time slice changes over time, and so our situation-liking can be expected to change with it. The problem with this proposal is that the things that matter to us take time—they do not happen in an instant. Thus our situation-liking is often affected by our beliefs about what is *going to happen*, and to a lesser extent by what has happened. Hence more than the current time slice is relevant.

We might accommodate the temporal variability of situation-liking by introducing the notion of a *temporally indexed possible world*, which formally is just an ordered pair consisting of a possible world and a time. As time passes, a person resides in the same possible world but

different temporally indexed possible worlds. However, this is really just a technical trick to manufacture an object for situation-liking. Even if this works, it is not clear that it illuminates anything. I think the better alternative is to deny that situation-liking is an intensional state. Situation-liking is really just a feeling of satisfaction. We might say that we are satisfied with "this time of our life", and so take the object of the satisfaction to be a time. However, nothing is gained by taking the time to be the literal *object* of the state. Reference to the time is just a way of saying that situation-liking is a state characterized by a parameter (the degree of situation-liking) that varies over time. In this sense it is like numerous other psychological states, including happiness, depression, fear, etc. We can be happy or depressed *about* something, and we can fear some particular thing, but we can also *feel* happy or depressed or afraid without those attaching mentally to any particular object. Thus although there is a sense of happiness, depression, and fear in which they are intensional states, there is also a sense in which they are nonintensional states. In the latter sense, they are *feelings*. In the same sense, situation-liking is a feeling.

The theory I am proposing is starting to look a lot like a traditional Benthamite theory of value. Jeremy Bentham proposed that the only intrinsic value is happiness. I am proposing that human evaluative cognition makes ultimate appeal to feelings of situation-liking. Another way to put this is that feelings of situation-liking are *treated cognitively as* the source of intrinsic value. My enterprise differs from Bentham's in that I am not trying to say *what is really of value*. Instead, I am trying to explain how evaluative cognition works. One could try to make an argument to the effect that if this is the way human evaluative cognition works, then situation-liking is the metaphysical source of intrinsic value. I am sympathetic to that line of reasoning, although I will not pursue it here.

I am framing my account of human evaluative cognition in terms of situation-liking, and I have said that situation-liking is a nonintensional state—a feeling, if you like. But just what psychological state is this? Bentham talked about feeling happy, and in discussing our foreign traveler I talked about her *enjoying* foreign travel. I also talked about feeling *satisfied* with one's current situation. Can situation-liking be identified with any of these familiar psychological states? Probably not. Situation-liking is more like a conglomerate of all the different kinds of positive feelings we can have. It is a kind of generic "pro attitude". We are all familiar with situation-liking through introspection, but most likely the only way to give a more precise description of it is "functionally"—by giving a general description of how the psychological state is used in cognition. The sense in which situation-liking is a pro attitude is that it provides the assessments of intrinsic values used in evaluating both goals and plans. So what my claim really amounts to at this point is that (1) there is a nonintensional psychological state that plays the functional role of grounding the assessments of intrinsic value used in evaluating goals and plans, and (2) humans have introspective access to this state. By the latter I do not mean that we have infallible access, any more than we have infallible access to other psychological states like beliefs and pains. The importance of this introspective access will be explored in sections seven and eight.

5.4 Correcting Situation-Liking

I have argued that situation-liking provides the basic premises for evaluative cognition, but this does not mean that it should always be taken at face value. A complication arises from the fact that our situation-liking is profoundly influenced by our beliefs. If we have false beliefs about our current situation, or lack some true beliefs, we may like it better or worse than we would if our beliefs were corrected. But in evaluating a prospective situation, if we would like being in that situation because we would have false beliefs about it, surely that does not make the situation desirable. For example, Nozick (1974) discusses the "experience machine". This is a machine that can make the world seem any way the agent chooses, and it can be set up so that the agent does not know he is in the experience machine once it begins operation. Suppose I

have some concrete goal, e.g., to construct a unified field theory for physics. In fact, the goal is beyond my ability. However, by subjecting myself to the experience machine I can make myself *think* that I have accomplished this goal, and that will make me quite content. Is this a reason for subjecting myself to the experience machine? Surely not. I want to *really* construct a unified field theory—not just *think* I have.

What this seems to indicate is that the evaluation of the agent's current situation should not be based simply upon the agent's current situation-liking, but rather counterfactually on what the agent's situation-liking would be if it were fully and accurately informed about its current situation. In trying to improve its situation, it is trying to render more likable the way the situation *actually is*—not just the way the situation is believed to be. The agent seeks to render its situation more likable, not just better liked. The objective is to change the situation so that it would be better liked *if* the agent had true beliefs about all the relevant aspects of it. In my (1995) I expressed this by saying that the agent seeks to improve the "objective likability" of its situation. Formulating this as a principle of evaluative cognition, it seems that liking one's current situation is a defeasible reason for attaching value to it, but discovering that one had false beliefs or lacked true beliefs that affected the situation-liking is a defeater for that reason and a reason for instead attaching a degree value reflecting what the situation-liking would have been had the agent been better informed. More comprehensive corrections of this sort take precedence over less comprehensive ones.

With this proviso, my suggestion is that we take situation-liking to provide the basic source of value in human evaluative cognition. I turn next to the task of constructing a more precise theory of evaluative cognition grounded upon situation-liking.

6. Situation-Based Evaluative Cognition

I argued in section three that computational feasibility requires the intrinsic values employed in selecting goals and evaluating plans to be derived from an evaluative database that constitutes a fundamental constituent of a rational agent's cognitive architecture. This general observation imposes no constraints on what the intrinsic values might be for any particular kind of agent. I suggest that in human beings feature-likings constitute the evaluative database.

In section four, I noted that there is a problem implementing the evaluative database in a sophisticated cognitive agent operating in an unconstrained environment. The concepts required for the construction of the evaluative database can only be acquired as a result of learning based upon extensive experience of the world. They cannot be built in from scratch. This suggests that the primitive elements of the evaluative database are not the basic source of values. Instead, they are derived from some more basic source.

In section five, I suggested that in human evaluative cognition, situation-liking provides the final court of appeal. Feature-likings can be computed by inquiring whether the liked situation-type has a high expected-value, where that expected-value is the mathematical expectation of the situation-liking of the agent when it is in situations of that type. The observations of section five were specifically about human cognition, but we can generalize them to construct a theory of *situation-based evaluative cognition* applicable to rational agents in general. This will, in turn, throw further light on human evaluative cognition.

6.1 Conative Dispositions

Situation-based evaluative cognition begins with situation-liking. An agent implementing such cognition must be equipped with a conative process that produces various degrees of situation-liking. The conative process proceeds in terms of *conative dispositions* to like one's current situation in response to various inputs. Several different conative dispositions may be

activated at one time, in response to different inputs, and their outputs are combined to form an overall situation-liking. What are the inputs to the conative dispositions? It seems clear that one kind of input is the agent's beliefs about its world and its situation in it. However, this produces an implementational problem that is analogous to the problem of implementing primitive feature-likings. Beliefs describe the world by taking it to exemplify concepts, so the conative dispositions that produce situation-liking in response to beliefs can equally be regarded as responding to concepts. As before, simple concepts might be built into the agent's cognitive architecture, but more complex concepts must be learned in conjunction with learning about the world. So there cannot be built-in conative dispositions making reference to complex concepts or complex beliefs about the agent's situation.

The human cognitive architecture solves this problem in part by including conative dispositions responsive to some nonconceptual inputs. Certain physiological states, like hunger, thirst, fatigue, sexual arousal, pain, etc., can affect situation-liking without the agent forming any beliefs. These "physiological" conative dispositions provide an initial source for situation-liking. Then as the agent begins to acquire beliefs about its world, some of those beliefs can acquire the ability to influence situation-liking through conditioning. In modern cognitive psychology, appeal to conditioning has fallen into disfavor. This is in response to its earlier overuse in trying to explain all of cognition. But here is a place where conditioning seems to be essential. There is no "more cognitive" mechanism that could create conative dispositions responsive to beliefs from conative dispositions responsive to physiological states, because by hypothesis the latter dispositions are nonconceptual. Conditioning (broadly construed) seems to provide the only possible mechanism for attaching value to beliefs without having the concepts built-in. However, it is a rather crude mechanism. The only way value gets attached to an abstract concept like democracy is via a long chain of causal connections in the world that is capable of supporting conditioning based ultimately on physiological sources of value.

Thus far I have argued that conative dispositions responsive to complex beliefs cannot be built in. They must be "acquired" in parallel with the agent learning the concepts involved in the beliefs. I have also argued that a kind of conative disposition that can be built in is a nondoxastic one responsive to physiological states. There remains a third possibility. Although complex concepts must be learned, simple ones might be built in, and correspondingly conative dispositions responsive to simple beliefs about the world might also be built in. Some such mechanism is presumably involved in a field mouse's innate fear of a hawk's shadow. However, the restriction to built-in concepts would seem to impose severe restrictions on the complexity of the beliefs. There is probably no way to build in fear of the hawk itself, but it is possible to build in fear of certain shapes because the concepts of those shapes can be innate rather than learned. Agents designed to work in only narrowly circumscribed environments can have a built-in world model and the concepts that go with it, and conation in such an agent can make use of a rich array of built-in concepts. But as we remove constraints, the epistemic cognition of the agent must become more sophisticated so that it can build its own world model in response to experience, and the resources available to conation become correspondingly more impoverished as the agent has fewer innate concepts for use in either a primitive evaluative database or built-in conative dispositions.

The argument against having built-in conative dispositions responsive to complex concepts is more compelling when applied to concepts pertaining to the external world than when applied to concepts pertaining to the cognitive agent itself. It is undesirable to build too much of an a priori theory of the external world into a cognitive agent, because that precludes its being able to deal with environments that conflict with its built-in theory. However, the same argument cannot be applied to building in an a priori theory of the agent itself. The structure of the agent is fixed and will be the same regardless of its external environment, so there is no obvious reason for not equipping the agent with a priori concepts descriptive of itself and other agents like itself. These concepts might be referenced by built-in conative dispositions, and this may play an

important role in the sociobiological aspects of human conation. Examples of the latter would include our natural conative responses to property, revenge, or various moral dimensions of our circumstances.

The combination of conative responses to physiological states and conative responses to beliefs involving simple (innate) concepts of the external world and possibly more complex concepts of ourselves seems like a rather blunt instrument for grounding the evaluative cognition of a rational agent, but it is hard to see what else could be involved. It becomes an open question how to design agents that will achieve specific design objectives using such crude conative tools. In particular, it remains to be seen how some of the more complex sociobiological aspects of human conation can be based on these crude mechanisms. If they can't be, then there must be some way to augment the present mechanism to accommodate them, but it is not at all clear how such an augmentation could be accomplished.

The above remarks were aimed primarily at human conation, but it is hard to see what alternative there could be for sophisticated cognitive agents operating in unconstrained environments. The only alternative would be to have situation-liking directly responsive to certain kinds of beliefs, but that requires the concepts involved in the beliefs to be built-in rather than learned, and that in turn requires either severe restrictions on the complexity of the beliefs or a built-in world model. For either artificial or natural agents operating in unconstrained environments as complex as the real world, there seem to be rather low limits on how much world knowledge can be innate.

6.2 Adopting Plans

I began this paper by describing cognitive agents as forming beliefs representing the world, evaluating the world as represented, forming plans for making the world more to their liking, and performing actions executing the plans. In other words, they implement the doxastic-conative loop. Situation-liking provides the evaluation of the world. Plans are adopted on the grounds that adopting and executing them can be expected to increase the agent's situation-liking over what it would otherwise be. The expectation involved in this evaluation of plans must take account of the fact that, in the real world, an agent cannot predict with certainty all of the contours of the situation that will result from executing the plan. There are numerous possible scenarios in which the plan is executed, each resulting in different situation-likings. The evaluation of a plan must proceed by evaluating all the different ways the world might be if the plan were executed, discounting each by the probability of its occurring, and then summing the results. This produces an *expected-value* for the plan. The details of the definition of "expected-value" for plans turn out to be surprisingly complex. See my (2001).

Plans cannot be evaluated in isolation. This turns upon the observation that sometimes changes to the agent's set of adopted plans should consist of adopting several plans and perhaps withdrawing other plans simultaneously. For instance, suppose an agent has a plan to visit a friend this afternoon. Then the need to run an errand arises. The value of the errand is high. Running the errand does not strictly preclude visiting the friend. It is still possible to visit the friend after running the errand, but that would make the visit too rushed to be enjoyable. Under the circumstances, the change the agent should make to its set of adopted plans is to simultaneously adopt the plan to run the errand and withdraw the plan to visit the friend. In other more complicated cases, the optimal change may be to adopt several plans and withdraw several plans simultaneously. So what should be evaluated decision-theoretically is not single plan adoptions, but rather changes to the agent's set of adopted plans, where the changes may consist of multiple additions and deletions. I will refer to these as *intention changes*.

6.3 Defining expected-values

The expected-value of plan (or more generally, an intention change) is to be understood

in terms of situation-liking. Notice, however, that executing a plan takes time, and our situation-liking can vary throughout the execution. Also, the goal that a plan is trying to achieve might be something that happens some time after the last step of the plan is executed. Evidently we cannot identify the expected-value of a plan with the expected-value of the situation-liking at any single instant. Executing a plan may result in substantial initial execution costs followed by a big gain in utility as the goal is achieved. To compute the overall costs and values achieved by executing the plan we must do something like integrating the situation-liking over time. This means that situation-liking is not treated as value, but rate of value production.

There may be numerous possible scenarios in which a plan is adopted and/or executed, each resulting in different time-courses of situation-likings. The value of a particular scenario is the result of integrating the situation-liking over the time.[12] What is the time interval over which the integration should be performed? There does not seem to be any natural cutoff. A scenario should not have a higher value associated with it just because it takes longer, so it seems that all scenarios should run for the life of the agent. (Of course, the agent can have different lifetimes in different scenarios.) Where *A* is an action, let us understand an *A-scenario* to be a temporal section of a possible world bounded temporally on the left by the time *A* is performed and on the right by the time the agent ceases to function. The actions we will be interested in will be intention changes. The *value of the scenario* is the result of integrating the agent's situation-liking over the length of the scenario. That is, where t_0 and t_1 are the temporal bounds of a scenario *S*:

$$\text{scenario-value}(S) = \int_{t_1}^{t_0} \text{situation-liking-at}(t)\ \mathrm{d}t$$

Note that the definition of "scenario-value" requires a cardinal measure of situation-liking. Otherwise, the integration makes no sense. Then at least as a first approximation,[13] the expected-value of an action is the mathematical expectation of the value of all possible scenarios in which the action is performed:

$$\text{expected-value}(A) = \int_{-\infty}^{\infty} r \cdot \frac{\mathrm{d}}{\mathrm{d}r}\, \text{PROB}(\text{scenario-value}(S) \leq r \ /\ S \text{ is an } A\text{-scenario})\ \mathrm{d}r$$

We will have occasion below to talk about the mathematical expectations of several functions, so let us abbreviate:

$$\text{EXP}(f(x)/\varphi x) = \int_{-\infty}^{\infty} r \cdot \frac{\mathrm{d}}{\mathrm{d}r}\, \text{PROB}(f(x) \leq r \ /\ \varphi x)\ \mathrm{d}r$$

Thus the expected-value of an action is given by EXP(scenario-value(*S*)/ *S* is an *A*-scenario).

6.4 Computing expected-values

The preceding discussion defines the expected-value of adopting a plan, but if agents are to use expected-values in deciding which plans to adopt, they must also have a way of computing them. The definition just given does not lend itself to direct implementation, because the set of

---

[12] Should future values be discounted? I see no reason to think so (other than limited life expectancy, which is already factored in when a probability is assigned to the scenario), but there is no theoretical obstacle to doing so.

[13] In my (2001) I argue that this definition must be made more complicated.

A-scenarios will normally be infinite, and each *A*-scenario, being a temporal-segment of a possible world, will be of infinite complexity. If expected-values are to be of any use in decision-making, the agent must have some finitistic way of computing (or estimating) them.

Representing expectations as integrals makes them hard to compute. However, if x ranges over only finitely many values $a_1,...,a_n$, the integral can be reduced to a finite sum (a weighted average of the $f(a_i)$'s):

$$\mathrm{EXP}(f(x) / \varphi x) = \sum_{0 \leq i \leq n} f(a_i) \cdot \mathrm{PROB}(x = a_i \ / \ \varphi x)$$

Attempts to make the computation of expectations feasible usually proceed by trying to reduce them to such finite sums. That will be the course taken here.

The first thing to notice is that in deciding what intention changes to make, the agent need not know the actual expected-values of the different changes. To choose an optimal change, the agent need only know how the expected-values compare with one another. For this purpose, it suffices to be able to compute the *differential expected-value*—the difference between the expected-value of a change and the expected-value of the unchanged set of adopted plans. This yields an important simplification to the decision problem. Most of the world will be left unchanged by an intention change, and those aspects of the world that are unaffected by the intention changes will probably be the principal contributors to situation-liking, but the agent can ignore the common aspects when evaluating the intention change, focusing just on what effect an intention change has on what the situation-liking would otherwise be.

When an agent considers a way the world might be as a result of adopting or executing a plan, it cannot conceive of an entire *A*-scenario. It must think of the world under a general description. The set of truths about the *A*-scenario is infinite, so the general description can be only a partial description, representing the agent's limited conception of the *A*-scenario. If the agent is to compute a value for the *A*-scenario, it must be computed on the basis of this general description. Human beings do this by selecting features of the description of the scenario that they expect to cause changes in their situation-liking, and assume that the situation-liking remains unchanged except for the changes wrought by changing those features. For example, imagine a scenario (scenario-type really) in which you are hungry, go to the kitchen and make a sandwich, and then eat it. In evaluating this scenario, you first imagine being hungry. Then you imagine going into the kitchen, which involves some slight effort (and corresponding negative value). Similarly, you imagine making the sandwich, attributing some small execution cost to that. Finally, you imagine eating the sandwich. You evaluate the latter both in terms of the pleasure you get from eating it and the alleviation of your hunger. In any actual world in which this scenario might be played out, infinitely many other things are happening at the same time but you ignore them. You assume that these are the only occurrences that change your situation-liking from what it would otherwise be. This manner of evaluating a scenario consists of picking out a small list of features that you expect to change your situation-liking, computing the values contributed by these changes (integrating the changes in the situation-liking over the time of the change), and summing the results.

It appears that this evaluation of features is precisely what is encoded in human feature-likings. We discover that certain features tend to cause certain cumulative changes in value (i.e., changes in situation-liking integrated over time). I will take the entries in the evaluative database to record these expected-values of state descriptions, where the expected-value of a state-description S is defined to be the mathematical expectation of the value caused by S:

$$\mathrm{EV}(S) = \mathrm{EXP}(\textit{value-caused-by}(S) \ / \ S \text{ occurs}).$$

Value-caused-by(S) is the cumulative situation-liking caused by S (i.e., the change in situation-liking

integrated over time). Using the evaluative database, we compute the overall changes wrought by a scenario by summing the changes wrought by particular states in the scenario. Those changes are computed using the database calculation of section three. That was represented as an organizational principle for the evaluative database rather than a substantive principle about values, but the reason it is a useful organizational principle is that we expect values to combine in that way except in unusual cases. The justification for this expectation will be discussed in the next section.

7. The Logical Credentials of the Reasoning

This section aims at making the preceding sketch of situation-based evaluative cognition logically precise. Thus far the proposal has four constituents:

- Plan adoption or retraction is performed as part of more general intention changes, which can result in adopting and retracting several plans at one time.
- Intention changes are compared in terms of their *differential expected-values*, i.e., the mathematical expectation of the change in situation-liking integrated over time.
- The differential expected-value of an intention change is computed by isolating a list of possible consequences of making the change that are *value-laden*, in the sense that they tend to cause changes in value.
  - Value-laden situation-types are stored in the agent's evaluative database.
  - The value of a possible scenario resulting from the intention change is computed using the database calculation discussed in section three.
- The differential expected-value of the intention change is then computed by taking a weighted average of the different possible scenarios thus generated, discounting each by its probability.

This proposal makes the computation of differential expected-values feasible by reducing it to the computation of a weighted average of value changes in finitely many scenario-types. The scenario-types are characterized by changes to value-laden states stored in the agent's evaluative database. This reasoning makes two assumptions. First, it assumes that the value assigned to a scenario-type can be computed by using the database calculation. Second, it assumes that the differential expected-value of an intention change can be computed in terms of values assigned to scenario-types characterized by changes to value-laden states. What are the logical credentials of these two assumptions?

7.1 The Evaluative Database

A scenario-type consists of a temporally ordered sequence of state descriptions. I propose to compute the expected-value of a scenario-type using the evaluative database. I will take the evaluative database to be organized according to the principles described in section three, with the caveat that the organization has to be relative to the cognitive agent's knowledge. That is, in section three it was supposed that the evaluative database represented the complete and correct assignment of values to state descriptions. Now the proposal is that the evaluative database represents the expected-values of state-descriptions *to the best of the agent's knowledge*. When the agent knows the expected-value of a conjunction and knows that it cannot be computed from the expected-values of its conjuncts in accordance with the database calculation (described in section three), then the true value is entered into the database as an explicit entry. If the value can be computed *or if the value is unknown*, then no entry is made in the database.

Organizing the database in this way has the consequence that the absence of an entry for a particular state description can mean either of two things. The value may be redundant, in the sense that it can be computed from other entries using the database calculation, or the value may be unknown. It may seem that this disjunctive significance makes the database useless for

retrieving values for items for which there are no explicit entries. If there is no entry for a state description because we have no independent knowledge of its value, but one can be computed using the database calculation, why should we expect that to be the right value? As I will now argue, the power of the evaluative database arises from the fact that the database calculation can be assumed defeasibly to return the right value even when we have no explicit knowledge of the value to be retrieved. This is what made the organizational principle efficient in the first place. If the database calculation did not normally produce the right answer, then all of the conjunctive state descriptions would have had to be entered into the database explicitly.

Recalling that $S_1 \cap S_2$ is the conjunction of the conjuncts S_1 and S_2 have in common, the database calculation can be grounded on two principles:

Irrelevance

If S is a state description and the agent does not know the expected-value of S or any subconjunction of S, it is defeasibly reasonable to take EV(S) to be 0.

Additivity

If $S_1,...,S_k$ are subconjunctions of a state description S, the agent knows the expected-values of each S_i but does not know the expected-value of S or of any subconjunctions subsuming any S_i, then it is defeasibly reasonable to take EV(S) to equal $EV(S_1)+...+EV(S_k) - \sum_{i\neq j} EV(S_i \cap S_j) + \sum_{i\neq j, i\neq k, j\neq k} EV(S_i \cap S_j \cap S_k) - ...$

But what reason is there to accept Irrelevance and Additivity? I will argue that these are grounded on principles that we commonly use for reasoning about causes. Irrelevance follows from the following general principle:

Causal Irrelevance

If the agent has no reason to think otherwise, it is defeasibly reasonable to think that P does not cause Q.

This is a general principle of causal reasoning that we employ regularly. We base our causal reasoning on those considerations that we know to be relevant, and we assume defeasibly that other considerations will not disrupt the causal connections that we know about. This has the consequence that if the agent has no reason to think otherwise, it is defeasibly reasonable to think that being in a situation described by a state description S does not cause a change in situation-liking, and hence EV(S) = 0.

To defend Additivity, we need the assumption that changes in situation-liking can be meaningfully combined. It must make sense to talk about "quantities" of situation-liking, so that the quantities can be combined to form larger quantities. One change can increase situation-liking, and then another change can produce "some more" situation-liking. In other words, "situation-liking" must be a mass noun. In this respect, situation-liking is like mass or area. It must also be assumed that we have a cardinal measure of situation-liking so that when quantities are combined, the size of the resulting quantity is the sum of the sizes of the constituent quantities. This has already been presupposed in talking about values caused by situation-types, where those values were defined to be the result of integrating changes in situation-liking over time. Such integration makes no sense without a cardinal measure of situation-liking. So I presume that we have a cardinal measure of situation-liking. This, however, is a big assumption. Where does this cardinal measure come from? I will try to answer this question in section eight. Given a cardinal measure of situation-liking, it follows that values (defined as cumulative situation-likings) are also quantities measured by a cardinal measure. Let us call such quantities *cardinal quantities.*

To justify additivity, we need to look more closely at the logical properties of cardinal quantities and cardinal measures. Cardinal quantities attach to "objects" (broadly construed).

For example, mass is the mass of a physical object or system of physical objects, and area is the area of a surface. Cardinal quantities are individuated by the objects to which they are attached. Talk of combining different "amounts" of a cardinal quantity is cashed out in terms of mereological operations on the objects to which the quantities attach. E.g., one and the same area cannot be the area of one surface at one time and a different surface at another time. There must be mereological operators $\oplus$ (union) and $\otimes$ (intersection) defined on the objects. For the areas of surfaces, $\oplus$ is set-theoretic union and $\otimes$ is set-theoretic intersection. For the masses of systems of objects, $\oplus$ merges two objects or systems of objects into a larger system, and $\otimes$ produces the overlap of two objects or systems of objects. The operators $\otimes$ and $\oplus$ are required to satisfy all the algebraic conditions satisfied by set-theoretic intersection and union (i.e., they constitute a semi-group). It is the attachment of cardinal quantities to objects and the ability to combine the objects mereologically that gives meaning to the idea that there are different "quantities" of the cardinal quantity and that they can be combined and manipulated mathematically. We can represent cardinal quantities as sets of "stuff" (mass, area, etc.). Letting $\varphi(x)$ be the cardinal quantity attached to an object x, unions and intersections of the cardinal quantity are understood in terms of mereological operations on the objects to which they attach:

$$\varphi(x) \cup \varphi(y) = \varphi(x \oplus y).$$

$$\varphi(x) \cap \varphi(y) = \varphi(x \otimes y).$$

A cardinal measure is generated by any additive set function f defined on the cardinal quantities. To say that f is additive is to say:

$$f(\mathrm{X} \cup \mathrm{Y}) = f(\mathrm{X}) + f(\mathrm{Y}) - f(\mathrm{X} \cap \mathrm{Y}).$$

A cardinal measure defined on the objects x is then:

$$F(x) = f(\varphi(x)).$$

It follows that

(1) $$F(x \oplus y) = F(x) + F(y) - F(x \otimes y).$$

Thus, for the example, the area of the union of two surfaces is the sum of the areas of the individual surfaces minus the area of their overlap. There can be different cardinal measures of the same cardinal quantity. For example, we can measure an area in cm^2 or in^2.

It would be natural to suppose that the mereology of objects to which values attach is the mereology of state descriptions defined by the operators "&" and "$\cap$". However, in accordance with principle (1), this would require that the following always be true:

(2) $$\textit{value-caused-by}(P\&Q) = \textit{value-caused-by}(P) + \textit{value-caused-by}(Q) - \textit{value-caused-by}(P \cap Q).$$

This fails in cases of value-theoretic interference, i.e., when the states are not value-theoretically independent. Thus the mereology of state descriptions cannot be used to give meaning to talk of quantities of value. However, there is another mereology that will do the trick. Values (cumulative situation-likings) are attached to state descriptions indirectly via causal processes whereby being in situations described by the state descriptions gives rise to some degree of situation-liking. Value-theoretic interference occurs when being in a situation described by one state description prevents the completion of the causal process ordinarily initiated by the second state description. So we can regard the values as being attached to the sets of causal processes that produce them,

and take the mereology to be the mereology of sets of causal processes. That is, if X and Y are sets of causal processes, X⊗Y will be X∪Y and X⊗Y will be X∩Y.

Quantities of value are individuated by the sets of causal processes that produce them, so where $v(X)$ is the quantity of value produced by a set of causal processes X, we automatically have:

(3) $v(X \cup Y) = v(X) \cup v(Y)$.

A cardinal measure *val* of value is then an additive measure defined on these quantities of value:

(4) $val(X \cup Y) = val(X) + val(Y) - val(X \cap Y)$.

Where S is a state description and X is the set of value-producing causal processes initiated by being in S, *value-caused-by*(S) is *val*(X). Although there is no logical guarantee that principle (2) holds, there is a defeasible reason for expecting it to hold normally. This turns on a generalization of the principle of causal irrelevance. Our causal knowledge is to the effect that certain causal processes operate in specific situation-types. However, there can always be more inclusive situation-types in which other events interfere with the causal processes. Because our knowledge of the current situation is always incomplete, we can never conclusively rule out the possibility of there being something that will interfere with causal processes that we expect, on the basis of our limited knowledge of the situation, to occur. Thus if we are to be able to draw any conclusions about the causal consequences of an event, we must assume defeasibly that if a causal process operates in one situation-type, it will continue to operate in a more inclusive situation-type. Let us symbolize "S_2 is a more inclusive situation-type than S_1" as "$S_1 \prec S_2$". So the principle is:

> "$S_1 \prec S_2$ and the causal process p operates (or does not operate) in S_1" is a defeasible reason for "the causal process p operates (or does not operate) in S_2".

Furthermore, knowledge of how causal processes operate in a more inclusive situation-type take precedence over how they operate in a less inclusive situation-type:

> If $S_1 \prec S_2 \prec S_3$ and different causal processes operate in S_1 than in S_2, then an inference to what causal processes operate in S_3 based upon knowledge of the causal processes operative in S_2 takes precedence over an inference based upon knowledge of the causal processes operative in S_1.

The ***principle of causal independence*** is the conjunction of these two principles of defeasible causal reasoning.

By the principle of causal independence, we can assume defeasibly that the causal processes operative in a situation described by a state description P will continue to operate in a situation described by the more inclusive state description (P&Q). Similarly, we can assume defeasibly that the causal processes operative in ($P \cap Q$) are those operative in both P and Q. Thus although there is no logical guarantee that principle (2) will hold, we can infer it defeasibly from principles (3) and (4). The full database calculation can be justified defeasibly in essentially the same way. This is forthcoming from a general principle about cardinal measures. We will often want to compute the cardinal measure of the result of joining several objects. Let us define the *additive measure* of a set of objects as follows:

$AF\{x_1,...,x_n\} = F(x_1 \oplus ... \oplus x_n)$.

We have the following theorem:

$$AF\{x_1,...,x_n\} = F(x_1) + AF\{x_2,...,x_n\} - AF\{x_1 \otimes x_2,...,x_1 \otimes x_n\}.$$

Proof:

$$\begin{aligned}
AF\{x_1,...,x_n\} &= F(x_1 \oplus ... \oplus x_n) \\
&= f(\varphi(x_1 \oplus ... \oplus x_n)) \\
&= f(\varphi(x_1) \cup ... \cup \varphi(x_n)) \\
&= f(\varphi(x_1) \cup (\varphi(x_2) \cup ... \cup \varphi(x_n))) \\
&= f(\varphi(x_1)) + f(\varphi(x_2) \cup ... \cup \varphi(x_n)) - f(\varphi(x_1) \cap (\varphi(x_2) \cup ... \cup \varphi(x_n))) \\
&= F(x_1) + f(\varphi(x_2) \cup ... \cup \varphi(x_n)) - f(\varphi(x_1) \cap (\varphi(x_2) \cup ... \cup \varphi(x_n))) \\
&= F(x_1) + f(\varphi(x_2 \oplus ... \oplus x_n)) - f(\varphi(x_1) \cap (\varphi(x_2) \cup ... \cup \varphi(x_n))) \\
&= F(x_1) + F(x_2 \oplus ... \oplus x_n) - f(\varphi(x_1) \cap (\varphi(x_2) \cup ... \cup \varphi(x_n))) \\
&= F(x_1) + AF\{x_2,...,x_n\} - f(\varphi(x_1) \cap (\varphi(x_2) \cup ... \cup \varphi(x_n))) \\
&= F(x_1) + AF\{x_2,...,x_n\} - f((\varphi(x_1) \cap \varphi(x_2)) \cup ... \cup (\varphi(x_1) \cap \varphi(x_n))) \\
&= F(x_1) + AF\{x_2,...,x_n\} - f((\varphi(x_1 \otimes x_2)) \cup ... \cup (\varphi(x_1 \otimes x_n))) \\
&= F(x_1) + AF\{x_2,...,x_n\} - f(\varphi((x_1 \otimes x_2) \oplus ... \oplus (x_1 \otimes x_n))) \\
&= F(x_1) + AF\{x_2,...,x_n\} - F((x_1 \otimes x_2) \oplus ... \oplus (x_1 \otimes x_n)) \\
&= F(x_1) + AF\{x_2,...,x_n\} - AF\{x_1 \otimes x_2,...,x_1 \otimes x_n\}.
\end{aligned}$$

Now let us apply this to values. Let $S_1,...,S_n$ be the unsubsumed primitively value-laden subconjunctions of some state description S. Define the *additive-value* of the set $\{S_1,...,S_n\}$ recursively as follows:

DEFINITION: $AV(\varnothing) = 0;$

$$AV\{S_1,...,S_n\} = \textit{value-caused-by}(S_1) + AV\{S_2,...,S_n\} - AV\{(S_1 \cap S_2),...,(S_1 \cap S_n)\}.$$

This is the value that $\{S_1,...,S_n\}$ would cause if its members operated independently. That is, where $X_1,...,X_n$ are the sets of value-producing causal processes initiated by being in $S_1,...,S_n$, $AV\{S_1,...,S_n\} = Aval\{X_1,...,X_n\}$. By the principle of causal independence, it is defeasibly reasonable to expect that $\textit{value-caused-by}(S) = AV\{S_1,...,S_n\}$. The justification for focusing on the unsubsumed primitively value-laden subconjunctions and ignoring those that are subsumed lies in the second part of the principle of causal independence, according to which an inference from more inclusive state descriptions takes precedence over an inference from less inclusive state descriptions.

The database calculation is now forthcoming. It follows from the definition of *additive-value* that

Theorem: $$AV\{S_1,...,S_n\} = \sum_{1 \le i \le n} \textit{value-caused-by}(S_i) - \sum_{i \ne j} \textit{value-caused-by}(S_i \cap S_j) + \sum_{i \ne j, i \ne k, j \ne k} \textit{value-caused-by}(S_i \cap S_j \cap S_k) - ...$$

This theorem is about values caused by states. It entails a principle about expected values. The mathematical expectation of a sum of functions is the sum of their mathematical expectations, so it follows that we can defeasibly expect:

$$EV(S) = EV(S_1) + ... + EV(S_k) - \sum_{i \ne j} EV(S_i \cap S_j) + \sum_{i \ne j, i \ne k, j \ne k} EV(S_i \cap S_j \cap S_k) - ...$$

This is the principle of Additivity.

My conclusion is that general principles about cardinal measures together with defeasible principles of causal reasoning make it defeasibly reasonable to expect the database calculation to

hold in any specific case. This justifies the use of the evaluative database in decision-theoretic reasoning.

7.2 Differential Expected-Values of Scenario-Types

The evaluative database is to be used in evaluating scenario-types. Recall that a scenario-type consists of a temporally ordered sequence of state descriptions, where each element of the sequence represents changes to value-laden states stored in the agent's evaluative database. The differential expected-value of a scenario-type is the mathematical expectation of the differential-values of scenarios of that type, and the differential-value of a scenario is the change in value from the scenario that would result from doing nothing. By the principle of causal irrelevance, it is defeasibly reasonable to expect changes in value to result only from changes in value-laden states recorded in the evaluative database. Thus it is defeasibly reasonable to identify the differential-value of a scenario with the sum of the changes in value wrought by the changes to members of the evaluative database. The mathematical expectation of a sum is the sum of the mathematical expectations, so it is defeasibly reasonable to identify the differential expected-value of the scenario-type defined by the sequence of changes in members of the evaluative database with the sum of the expected-values of the state-descriptions comprising the scenario type. We thus have a computationally feasible way of computing differential expected-values for scenario-types.

7.3 Evaluating Intention Changes

Intention changes (i.e., changes to the set of adopted plans) are evaluated in terms of their differential-expected-values, which were defined as follows:

> differential-expected-value(*A*)
> = EXP(differential-expected-value(*T*) / *T* is an *A*-scenario-type).

The nature of this calculation depends upon the structure of the adopted plans. Consider an intention change that consists of adopting a simple linear plan (i.e., a plan prescribing a linearly ordered sequence of actions), and suppose the agent has a number of beliefs to the effect that performing those actions under various circumstances will, with certain probabilities, result in changes to particular value-laden states. By the principle of causal irrelevance, the agent can assume defeasibly that there are no unknown causes of changes to the value-laden states. The agent can then construct a tree of state descriptions linked by probabilities, analogous to the trees constructed by Markov decision planners (but with a much smaller set of states). Each branch through the tree represents an *A*-scenario-type, and the tree provides the information needed to compute a probability for that *A*-scenario-type. The differential-expected-value of adopting the plan is then the sum of the values of the finitely many scenario-types represented in the tree, each discounted by its probability.

Plans with more complex structures will generate trees with more branches. For example, a plan prescribing a partially ordered sequence of actions can produce different scenario-types depending upon the order in which the plan steps are executed. But the general computation of expected-values is the same.[14]

The decision whether to make an intention change is then based upon a comparison of its differential-expected-value with that of the "null change", which consists of retaining the current set of adopted plans. The differential-expected-value of the null change is automatically 0.

---

[14] See my (2001) for more details.

7.4 Building a Cognitive Agent

The preceding provides a logically precise characterization of "situation-based evaluative cognition". My main claim is that it should be possible to build a cognitive agent that works this way. My secondary claim is that it is hard to see how to build a cognitive agent capable of functioning in unconstrained environments that works in any other way.

Situation-based evaluative cognition requires some crucial epistemological capabilities. First, for an agent to work in this way, it must be able to tell how much it likes its current situation. In other words, it must have epistemic access to its degree of situation-liking. All of the value-theoretic concepts employed in situation-based evaluative cognition are defined ultimately in terms of a cardinal measure of situation-liking, so if the agent is to be able to compute expected-values, it must begin by knowing the values of that cardinal measure. In building an artificial agent, that is unproblematic. It is computationally trivial to give the agent the ability to introspect those values. But for humans, this is obviously a problem, because we cannot introspect numbers that measure our situation-likings. I will return to this issue in the next section.

The second epistemological requirement is that the agent be able to acquire the requisite causal and probabilistic knowledge. Given sufficient epistemic sophistication, a cognitive agent should be able to do this. The only question is whether it can do it quickly enough. If we reflect upon the learning of human infants, it looks like the task is not insurmountable. For example, a child can learn quickly that putting its finger in a candle flame causes pain. On the other hand, without the protection of adults a child could not survive in a hostile environment long enough to acquire such causal knowledge. The same is probably true of artificial cognitive agents. It will take them awhile to acquire enough causal knowledge to be able to make their way in the world, so they too will have to be protected and tutored in their intellectual infancy.

8. Analogue Representations of Values

Thus far I have sketched a theory of situation-based evaluative cognition, and suggested that it represents the only way evaluative cognition can work in sophisticated cognitive agents capable of functioning in unconstrained environments. But there is a problem with this claim. An essential feature of situation-based evaluative cognition as I have formulated it is that agents are able to introspect their degrees of situation-liking and manipulate them mathematically in computing expected values. That would seem to require them to introspect numerical values for situation-liking. Human beings are not able to do that, but they are quintessential general-purpose cognitive agents. How can these two observations be reconciled?

The observation that humans are not able to introspect numerical values for their conative states is responsible for contemporary rational choice theory's retreat to preference rankings in place of numerical measures of value. However, I argued in section two that preference rankings cannot provide the computational basis for evaluative cognition. This is not to deny that humans have preferences, but just to deny that the preferences constitute the basic evaluative database from which evaluative cognition proceeds. The only computationally feasible way of grounding evaluative cognition in an agent capable of functioning in a complex environment is via a cardinal measure of value.

8.1 Analogue Representations of Quantities

The solution to this quandary lies in recognizing that numbers are not the only possible mental representations of quantities. Humans regularly employ and manipulate what are sometimes called "analogue representations" of quantities. For example, consider human judgements about length. We are quite good at comparing lines and judging which is longer. This might suggest that all we are really able to do perceptually is make an ordinal comparison,

judging that one line is longer than another. This would be analogous to the claim that our fundamental access to values is via preferences. But in fact, we are able to make more elaborate comparisons of lengths. We can judge not only that one line is longer than another, but also that it is *much longer*, or perhaps *just a little longer*. Such judgments make no sense if we have only ordinal comparisons. We can even judge perceptually that one line is *twice as long* as another. For that to make sense, we must be employing something like a cardinal measure, even though we are not using numbers to represent the sizes. Instead we employ analogue representations. Analogue representations can be mapped onto numbers, and the numbers behave like a cardinal measure. It is common in psychometrics to construct a rough mapping of this sort by asking subjects to rate quantities on a scale of 1 to 10. This is often something that subjects are able to do fairly easily and with considerable consistency.

The ability to compare lengths perceptually enables us to construct rulers and introduce numerical measures of length, and as a result of making a mental comparison of a line with a remembered ruler one is often able to look at a line and make a numerical estimate of its length. But it is important to realize that the resulting number is not our primary representation of the length. We can judge that one line is twice as long as another without first constructing numerical estimates of their lengths. In fact, the ability to judge, for instance, that a line is 1 1/2 feet long, trades on a prior ability to judge that it is 1 1/2 times as long as a one foot ruler. Humans naturally employ analogue representations of quantities and construct numerical representations artificially by comparing the analogue representation of a length with the analogue representation of the length of a standard unit of measure.

8.2 Q&I Modules

Humans not only employ analogue representations of length—they also manipulate them mathematically. For example, we can judge that one length is the sum of two others. This is the kind of mathematical operation that normally requires a cardinal measure. We might say that human analogue representations of lengths constitute a kind of nonnumerical cardinal measure, because associated with those analogue representations are mental operations that correspond to addition.

Elsewhere (Pollock 1989, 1995) I introduced the notion of a *Q&I ("quick and inflexible") module*. Q&I modules are fast special-purpose cognitive modules that supplement reasoning. For example, in catching a baseball we rely upon a Q&I module that allows us to quickly predict trajectories. If we had to predict the path of a flying baseball by reasoning about parabolic trajectories, the ball would long since have passed us by before we would be in a position to try and catch it. Q&I modules are fast, but they are also inflexible in the sense that they achieve their speed in part by making assumptions about the environment. For example, our trajectory module works on the assumption that the path of the ball will be unobstructed. If the ball is going to hit a tree, we must wait until we see its new path before we can predict its trajectory.

It is significant that the prediction of trajectories employs analogue representations of speed and direction. We could solve the same prediction problem by explicit reasoning (albeit much more slowly), but first we would have to have numerical measurements of speed and direction. We do not normally have such numerical information at our disposal.

Human cognition employs a number of different Q&I modules. For example, most of our reasoning about probabilities seems to be based upon analogue representations of probabilities and their manipulation by Q&I modules. It is rare for people to have actual numerical values for the probabilities they cognize about in everyday life. We can attempt to construct numerical measures by comparing our analogue representations of probabilities with analogue representations of probabilities in games of chance for which numerical measures are readily available. Subjective probability theory is based upon such an approach. Note that this is essentially similar to the way in which we construct numerical measures of lengths by comparison with

rulers.[15]

8.3 Numbers

It seems fairly clear that our normal representation of quantities is analogue rather than numerical, and our normal way of manipulating analogue representations is by employing various kinds of Q&I modules. The numerical representation of the cardinality of a set may be a built-in mode of representation in the human cognitive architecture, but its extension to the numerical representation of other kinds of quantities appears to be a human discovery or invention rather than a built-in mode of representation. We learn how to assign numbers to other kinds of quantities by discovering how to compare the quantities with standardized quantities that have numbers associated with them in some natural way. Our manipulation of analogue quantities then corresponds to the mathematical manipulation of numbers. Some Q&I modules correspond to very sophisticated mathematical computations. For instance, humans can reason about areas in much the same way they reason about lengths. Suppose, for example, that I draw three irregular figures on the board and ask you to draw a figure whose area is the average of the areas of the given figures. The result might be something like figure two. It is extremely interesting that this is a task we can actually perform, and without great difficulty. The results are unlikely to be exactly right, but they will be approximately right. To perform this same task by measuring the areas and constructing a new figure with the average area of the three given figures would be extraordinarily difficult.

Figure 2. Averaging Areas.

[15] However, it is of some importance that what this produces is a numerical representation for our *estimate* of the probability—not necessarily the true value. Subjective probability theory tries instead to *define* probability as the result of such a comparison. That approach is fraught with difficulty. See the discussion of subjective probability in Pollock and Cruz (1999), 93-98.

8.4 Situation-Liking and Value

Having surveyed some other uses of analogue representations for quantities and their manipulation by Q&I modules, it is a small step to the conclusion that human evaluative cognition employs similar tools. I remarked above that humans can introspect not only that they like one thing better than another, but also that they like it a lot better, or perhaps just a little better. Such comparisons make no sense if we suppose that humans are capable of only ordinal comparisons. It seems undeniable that when we make such judgments we are employing analogue representations of value-theoretic quantities.

If human evaluative cognition is to be an implementation of situation-based evaluative cognition, then humans must be able to compute a variety of kinds of expected-values. We require expected-values for scenario-types, scenarios, state descriptions, and intention changes. These are defined to be the results of complex calculations in the integral calculus. Section seven showed how to reduce these to finite sums of products of probabilities and values, but the mathematics required for actual cases can still seem formidable. However, it is no more formidable than the mathematics required for averaging the areas of irregular figures, and humans can do that quickly and efficiently. I suggest that we are likewise equipped with Q&I modules that enable us to compute expected-values with equal ease (although again, with less than total precision). For example, how do we compute (an analogue representation of) the expected-value of a scenario-type? We *imagine it*, ask ourselves how much we would like or dislike it, introspect the result, and take that as our analogue representation of the estimated expected-value. Similarly, in computing an expected-value for a plan, we consider the different scenario-types that might result from executing it, consider how probable they are (using analogue representations of probabilities), and then "mush them all together" to produce an analogue representation of the weighted-average which is the expected-value. The operation of "mushing them all together" is a Q&I module whose function is precisely that of computing expected-values.

The upshot of this is that humans really can perform the cognitive tasks required for situation-based evaluative cognition. What they *cannot* do easily is convert the tasks into explicit mathematics and solve the mathematical problems, but that is not necessary. Q&I modules operating on analogue representations solve the same problems, just as they do almost anywhere that humans deal with quantities.

In talking about values, it is customary to measure them in terms of "utiles". This can seem puzzling, because no one has ever proposed a numerical scale of utiles or explained how to actually measure values in terms of utiles. This becomes less puzzling when we realize that humans employ analogue representations of value in evaluative cognition. No scale problem arises for analogue representations. However, the manipulation of analogue representations of cardinal quantities is generally rather imprecise. For other kinds of cardinal quantities, like length or area, we increase the precision of our reasoning by discovering ways of using numbers in place of analogue representations. That is done by devising ways of measuring cardinal quantities by comparing them with standard units of the quantity. For example, we measure lengths using rulers. It is important to realize that this bootstraps on a prior ability to manipulate analogue representations of lengths. We cannot use a ruler for measuring lengths unless we know that its length does not change when we move it around. You cannot make rulers out of silly putty. And you cannot determine that your rulers are rigid by measuring them with rulers. That leads to an infinite regress. So the ability to judge lengths using rulers presupposes a prior ability to judge lengths non-numerically.

We could likewise improve the precision of "scientific" evaluative cognition if we could discover a way of measuring values numerically. Just as in the case of lengths, this would bootstrap on a prior ability to employ analogue representations of value. Perhaps this can be

done somewhat in the manner proposed by Ramsey (1926) and von Neumann and Morgenstern (1944), in terms of preferences between lotteries. But the details remain to be worked out.

9. Conclusions

I have argued on largely computational grounds that evaluative cognition in sophisticated cognitive agents capable of functioning in unconstrained environments must accord with the model of situation-based evaluative cognition. To briefly recapitulate, cognitive agents implement the doxastic-conative loop. The feasible implementation of the doxastic-conative loop in a complex environment requires goal-directed planning and the decision-theoretic evaluation of plans. The latter must proceed in terms of an evaluative database assigning a cardinal measure of values to concepts that pick out value-laden situation-types. In an epistemically sophisticated agent capable of operating in unconstrained environments, most of the concepts assigned values in the evaluative database must be learned rather than built in. Thus the evaluative database cannot be built in either. It must be derived from something else. The only obvious candidate for a source for the evaluative database is situation-likings produced by conative dispositions responsive to non-doxastic inputs and beliefs employing only simple concepts. The values stored in the evaluative database are then expectations of situation-liking integrated over time. In human beings, the computations required for this schema of situation-based evaluative cognition are implemented in terms of analogue representations of quantities and Q&I modules that perform mathematical operations on the analogue representations.

References

Boutilier, Craig, Thomas Dean, Steve Hanks
1999 "Decision-theoretic planning: structural assumptions and computational leverage", *Journal of Artificial Intelligence Research* **11**, 1-94.

Landauer, Thomas K.
1986 "How much do people remember? Some estimates of the quantity of learned information in long-term memory", *Cognitive Science* **10**, 477-494.

Millgram, Elijah
1997 *Practical Induction*, Cambridge, Mass: Harvard University Press.

Nozick, Robert
1974 *Anarchy, State, and Utopia*, New York: Basic Books.

Pollock, John L.
1993 "The phylogeny of rationality", *Cognitive Science* **17**, 563-588.
1995 *Cognitive Carpentry*, Cambridge, Mass: Bradford / MIT Press.
1998 "The logical foundations of goal-regression planning in autonomous agents", *Artificial Intelligence* **106**, 267-335.
2001 "Logical foundations for decision-theoretic planning in autonomous agents". In preparation. Available from http:/ / www.u.arizona.edu / ~pollock.

Pollock, John L., and Joseph H. Cruz
1999 *Contemporary Theories of Knowledge, second edition*, Lanham, MD: Rowman and Littlefield.

Ramsey, Frank
1926 Truth and probability. In *The Foundations of Mathematics*, ed. R. B. Braithwaite. Paterson, NJ: Littlefield, Adams.

Savage, Leonard
1972 *The Foundations of Statistics.* 2nd Edition. New York: Dover.

von Neumann, J., and Morgenstern, O.
1944 *Theory of Games and Economic Behavior*. New York: Wiley.