Discussion Paper No. 1127

CASE-BASED KNOWLEDGE and PLANNING*

by

Itzhak Gilboa+

and

David Schmeidler++

April 1995

* We are grateful to Eva Gilboa, Kimberly Katz, Akihiko Matsui, and Ariel Rubinstein for discussions and references. This paper supersedes "Case-Based Knowledge Representation: An Expository Note." A reader who is familiar with the previous paper is advised to read only Sections 5 and 6, which are new.

+ KGSM-MEDS, Northwestern University, Leverone Hall, Evanston, IL 60208.

++ Department of Statistics and Recanati School of Business, Tel Aviv University, Tel Aviv 69978, Israel; Department of Economics, Ohio State University, Columbus, OH 43210-1172. Currently visiting the Department of Economics and KGSM, Northwestern University, Evanston, IL 60208.

Abstract

"Case-Based Decision Theory" is a theory of decision making under uncertainty, suggesting that people tend to choose acts that performed well in similar cases they recall. The theory has been developed from a decision-/game-/economic-theoretical point of view, as a potential alternative to expected utility theory. In this paper we attempt to re-consider CBDT as a theory of knowledge representation and of planning, to contrast it with the rule-based approach, and to study its implications regarding the process of induction.

1. Introduction

The dominant paradigm in decision, game and economic theory for decision making under uncertainty is expected utility theory (EUT). It suggests that a decision maker can be ascribed a utility function and a subjective probability measure, so that her decision can be described by maximization of the expected utility. While there is no doubt that this is an elegant an powerful theory, we find it cognitively implausible in many decision situations, and in Gilboa and Schmeidler (1992) we offer an alternative approach, "Case-Based Decision Theory" (CBDT). It suggests that people tend to choose acts that performed well in similar cases they recall. In this paper we attempt to re-consider CBDT as a theory of knowledge representation and of planning, to contrast it with the rule-based approach, and to study its implications regarding the process of induction.

We start with an overview of CBDT (Section 2), followed by a brief account of the motivation for its development (Section 3). We then proceed to compare it to the rule-based approach, arguing that the implicit induction performed by case-based decision makers avoids some of the pitfalls of explicit induction (Section 4). In Section 5 we extend CBDT to deal with planning, and provide an axiomatic derivation of the suggested planning procedure. Section 6 is devoted to the process of induction from a case-based perspective, and also helps to delineate the boundaries of CBDT as presented earlier. We first discuss two levels of induction, and argue that "second-order" induction might call for generalizations of CBDT in its present (linear) form. We then contrast two views of induction, one based on simplicity, the other – on similarity. The comparison of these offers additional examples in which linear CBDT might be too restrictive to capture inductive reasoning. Finally, Section 7 concludes.

2. An Overview of CBDT

Case-Based Decision Theory views cases as instances of decision making. It therefore splits each "case" to three components: the decision problem, the act that was chosen in it by the decision maker, and the outcome she has experienced. Formally, we postulate three abstract sets as primitive:

P – the set of (decision) *problems*

A – the set of possible *acts*

R – the set of conceivable *results*.

and define the set of *cases* to be the set of ordered triples, or the product of the above:

$$C \equiv P \times A \times R \ .$$

At any given time, the decision maker has a *memory*, which is a finite subset of cases $M \subseteq C$.

We argue that decisions are based on similar cases in the past. In the basic model, similarity is a relation between decision problems (rather than whole cases). Desirability judgments, on the other hand, are assumed to solely depend on the cases' outcomes. We thus postulate a *similarity function* which may be normalized to take values not exceeding unity:

$$s : P^2 \to [0,1]$$

and a *utility function*

$$u : R \to \Re \ .$$

CBDT prescribes that acts be evaluated by a similarity-weighted sum of the utility they yielded in past cases. That is, given a memory $M \subseteq C$ and a problem $p \in P$, every act $a \in A$ is evaluated by the functional

$$U(a) = U_{p,M}(a) = \sum_{(q,a,r) \in M} s(p,q)u(r)$$

where a maximizer of this functional is to be chosen. (In case the summation is over the empty set, the act is assigned a "default value" of zero.)

Viewing CBDT as a descriptive theory, that supposedly describes how people make decisions, one wonders, is it refutable? Are there any modes of behavior that cannot be accounted for by an appropriately defined similarity function? To what extent are the theoretical concepts of "similarity" and "utility" observable? In Gilboa and Schmeidler (1992) we provide an axiomatic derivation of the similarity function, coupled with the decision rule given above. That is, we assume as datum a decision maker's preference order over acts, given various conceivable memories. We impose certain axioms on this order, which are equivalent to the existence of an essentially-unique similarity function, such that the maximization of U (using this similarity function) represents the given order. (In our derivation the utility function is assumed known. However, similar, though less elegant axioms would give rise to a simultaneous derivation of the utility and the similarity functions, in the context of U-maximization.)

As many other theories in the social sciences, CBDT should only be taken as a "first approximation," rather than an accurate description of reality. Further, there is little doubt that it may be more appropriate for certain applications, and less for others. In particular, there are two variations on the basic theme that are relevant to the sequel. The first is the "averaged similarity" version; the second – the "act similarity" generalization. We describe them below.

The CBDT functional U is cumulative in nature. The impact of past cases is summed up; consequently, the number of times a certain act was chosen in the past affects its perceived desirability. For example, consider a problem with a memory where all similarity values are 1, and where act a was chosen ten times, yielding the utility 1 in each case. Compare it to act b which was chosen twice, yielding a utility value 4. U maximization would opt for a over b. By contrast, it makes sense to consider a similarity-based "average" utility, namely

$$V(a) = \sum_{(q,a,r)\in M} s'(p,q)u(r)$$

where

$$s'(p,q) = \begin{cases} \dfrac{s(p,q)}{\sum_{(q',a,r)\in M} s(p,q)} & \textit{if well-defined} \\ 0 & \textit{otherwise} \end{cases} .$$

In Gilboa and Schmeidler (1992) we also provide an axiomatic derivation of V-maximization. In the interpretation of the functional V, memory serves only as a source of information regarding the performance of various acts. By contrast, in the interpretation of U memory can also be interpreted as affecting preferences directly: the accumulation of positive utility values reflects "habituation," while that of negative ones – "cumulative dissatisfaction" or "boredom aversion."

Both functionals U and V judge an act's desirability based on its own history. In many situations, it appears likely that past performance of other, similar acts will color the perception of a given act. That is, the similarity function might be extended to problem-act pairs, yielding the following functional:

$$U'(a) = U'_{p,M}(a) = \sum_{(q,b,r)\in M} s((p,a),(q,b))u(r) \ .$$

In Gilboa and Schmeidler (1994) we axiomatize this decision rule (again, assuming as given the concept of "utility"). One may combine these two variations and consider "averaged problem-act-similarity," in which the similarity values above are normalized so that they sum to 1 (or zero) for each act. Let V' denote the corresponding evaluation functional.

For the purposes of the ensuing discussion, it might be convenient to think of a further generalization, in which the similarity function is defined over cases, rather than problem-act pairs. According to this view, the decision maker may realize that a similar act in a similar problem may lead to a correspondingly *similar* (rather than identical) result. For instance, assume that our decision maker is buying a product in a store. In the past, different prices were posted by the product. Every time she decided "to buy," the result was having the product but parting from the posted amount of money. The decision maker is now in the store, facing a new price. We would expect her to imagine, based on her experience, that the buying decision would result in an outcome in which she has less money than when she entered the store, and that the difference be the new price. While one may attempt to fit this type of reasoning into the framework of U'-maximization by a re-definition of the results, it is probably most natural to assume a similarity function that is defined over whole cases. Thus, the case "the price is $10, I buy, and have the product but $10 less" is similar to the case "the price is $12, I buy, and have the product but $12 less." If we assume that the decision maker can imagine the utility of every outcome (even if it has not been actually experienced in the past), we are naturally led to the following generalization of CBDT:

$$U''(a) = U''_{p,M}(a) = \sum_{(q,b,t)\in M} s((p,a,r),(q,b,t))u(t) \ .$$

We do not provide an axiomatic derivation of U''-maximization. However, we will include both this rule and the corresponding (averaged) V''-maximization in the general class of linear CBDT functionals.

3. CBDT and EUT

Expected utility theory (EUT) suggests that people behave as if they were maximizing the expectation of a utility function based on some subjective probability measure. The expected utility model assumes a space of "states of the world," each of which "resolves all uncertainty" (as stated by Savage (1954)), describing the outcome of every act the decision maker might take. EUT is a powerful and remarkably elegant theory. Further, there is no denial that it is very useful both as a descriptive and as a normative theory of decision making. However, we claim that it is useful mostly when it is cognitively plausible, and that it loses much of its appeal when the notion of "state of the world" becomes a vague theoretical construct that is not "naturally" given in the description of the decision problem. For such problems we suggest CBDT as an alternative. Much of Gilboa and Schmeidler (1992) is devoted to comparisons of EUT and CBDT. Here we only mention a few relevant points.

The very description of a "decision problem" in EUT requires some hypothetical reasoning; should the decision maker reason in terms of EUT, she would have to imagine what would be the outcome of each act at each state of the world. Then she would have to assess the utility of each conceivable outcome, and the probability of each state. We argue that, unless the problem has been frequently encountered in the past, there is no basis for the assessment of probabilities (the "prior"). Moreover, imagining all relevant states is often a daunting cognitive task in itself. Correspondingly, in such situations people are likely to violate the seemingly-compelling Savage axioms (which give rise to expected utility maximization).

By contrast, CBDT requires no hypothetical reasoning: the decision maker is assumed to know only those cases she has witnessed, and to assess similarity values only for the problems she has encountered. Furthermore, in the original version of the theory the decision maker is not even assumed to "know" her own utility – she only needs to judge the desirability of the outcomes she has actually experienced. (Admittedly, this last principle is compromised in the generalized versions of CBDT described above as U''- or V''-maximization.)

In CBDT, an act that has not been tried before is assigned the default value of zero. This value may be interpreted as the decision maker's "aspiration level:" as long as it is obtained, the decision maker is "satisficed" (à la Simon (1957) and March and Simon (1958)) and will keep choosing the same act; if it is not obtained, she will be "dissatisficed" and will be prodded to experiment new acts.

In EUT the decision maker is implicitly assumed to be born with beliefs about everything she might encounter, and she learns by excluding things that *can not* happen (and updating the probabilities by Bayes' rule). By contrast, a case-based decision maker knows nothing at the outset, and learns primarily by adding cases to memory, that is, by learning what *can* happen.

EUT seems to be well-suited to problems that recur in more-or-less the same form, thereby allowing the decision maker to realize what the possible states of the world are, what their relative frequencies are, and so forth. The main appeal of CBDT as a descriptive theory is in those cases in which states of the world are not naturally given. However, it also can be viewed as a description of how people learn to be EU maximizers: the accumulation of cases in memory is a way to learn what the possible eventualities are, what their likelihood is, and so forth. Furthermore, with appropriate assumptions on the way in which the aspiration level is updated, case-based decision makers can be shown to converge to choosing EU-maximizing alternatives, provided they are actually faced with the same problem repeatedly (Gilboa and Schmeidler (1993)).

4. CBDT and Rule-Based Knowledge Representation

4.1 What Can Be Known?

Much of the literature in philosophy and artificial intelligence (AI) assumes that one type of objects of knowledge are "rules," namely general propositions of the form "For all x, $P(x)$." While some of these rules may be considered, at least as a first approximation, "analytic propositions," a vast part of our "knowledge" consists of "synthetic propositions."[1] These are obtained by induction, that is, by generalizing particular instances of them.

The process of induction is very natural. Asked, "What do you know about...?", people tend to formulate rules as answers. Yet, as was already pointed out by Hume, induction has no logical justification. He writes (Hume 1748, Section IV),

[1] By "synthetic" propositions we refer to non-tautological ones. While this distinction was already rejected by Quine (1953), we still find it useful for the purposes of the present discussion.

> "... The contrary of every matter of fact is still possible; because it can never imply a contradiction, and is conceived by the mind with the same facility and distinctness, as if ever so conformable to reality. *That the sun will not rise to-morrow* is no less intelligible a proposition, and implies no more contradiction than the affirmation, *that it will rise.* We should in vain, therefore, attempt to demonstrate its falsehood."

That is, no synthetic proposition whose truth has not yet been observed is to be deemed necessary. In particular, useful rules – that is, rules that generalize our experience and have implications regarding the future – cannot be *known*.[2]

Since induction may well lead to erroneous conclusions, it raises the problem of knowledge revision and update. Much attention has been devoted to this problem in the recent literature in philosophy and AI. (See Levi (1980), McDermott and Doyle (1980), Reiter (1980) and others.) In the spirit of Hume, it is natural to consider an alternative approach that, instead of dealing with the problems induction poses, will attempt to avoid induction in the first place. According to this approach, knowledge representation should confine itself to those things that can indeed be known. And these include only facts that were observed, not "laws;" cases can be known, while rules can at best be conjectured. One of the theoretical advantages of this approach is that, while rules tend to give rise to inconsistencies, cases cannot contradict each other.

Our approach is closely related to (and partly inspired by) the theory of Case-Based Reasoning (CBR) proposed by Schank (1986) and Riesbeck and Schank (1989). (See also Kolodner and Riesbeck (1986) and Kolodner (1988).) In this literature, CBR is proposed as a better AI technology, and a more realistic descriptive theory of human reasoning than rule-based models (or systems). However, our approach differs from theirs in motivation, emphasis and the analysis that follows. We suggest the case-based approach as a solution to, or rather a way to avoid the theoretical problems entailed by explicit induction. Our focus is decision-theoretic, and therefore our definition of a "case" highlights the aspect of decision making. Finally, our emphasis is on a formal model of case-based decisions, and the extent to which such a model captures basic intuition.

One has to admit, however, that even cases may not be "objectively known." The meaning of "empirical knowledge" and "knowledge of a fact" are also a matter

[2] Quine (1969) writes, "... I do not see that we are further along today than where Hume left us. The Humean predicament is the human predicament."

of heated debate. (For a recent anthology on this subject, see Moser (1986).) Furthermore, as has been argued by Hanson (1958), observations tend to be theory-laden; hence the very formulation of the "cases" that we allegedly observe may depend on the "rules" that we believe to apply. It therefore appears that one cannot actually distinguish cases from rules, and even if one could, such a distinction would be to no avail, since cases cannot be known with certainty any more than rules can.

While we are sympathetic to both claims, we tend to view them as somewhat peripheral issues. We still believe that the theoretical literature on epistemology and knowledge representation may benefit from drawing a distinction between "theory" and "observations," and between the knowledge of a case and that of a rule. Philosophy, like other social sciences, often has to make do with models that are "approximations," "metaphors," or "images," – in short, models that should not be expected to provide a complete and accurate description of reality. We will therefore allow ourselves the idealizations according to which cases can be "known," and can be observed independently of rules or "theories."

4.2 Case-Based Decision Theory

One may agree that only cases can be known, and yet question the appropriateness of the CBDT model. Indeed, many other representations of cases are possible, and one may suggest many alternatives to the CBDT functionals described above. We now turn to justify the language in which CBDT is formulated, as well as the basic idea underlying the linear functionals, namely, the similarity-weighted aggregation of cases.

We take the view that knowledge representation is required to facilitate the use we make of the knowledge. And we use knowledge when we act, that is, when we have to make decisions. This implies that the structure of a "case" should reflect the decision-making aspect of it.

Let us first consider cases that involve decisions. Focusing on the decision made in a given case, it is natural to divide a "case" to three components: (i) the conditions at which the decision was made; (ii) the decision; and (iii) the results. If we were to set our clocks to the time of decision, these components would correspond to the past, the present and the future, respectively, as perceived by the decision maker. In other words, the "past" contains all that was known at the time of decision, the "present" – the decision itself, while the "future" – whatever followed from the "past" and the "present." In the model presented above we dub the three components "problem," "act," and "result."

There are, indeed, cases that are relevant to decision making, without involving a decision themselves. For instance, one might know that clouds are typically followed by rain. That is, one has observed many cases in which there were clouds in the sky, and in which it later rained. These cases would fit into our framework by suppressing the "act," and re-interpreting "a problem" as "circumstances."

How are the cases used in decision making, then? Again, we resort to Hume (1748) who writes,

> "In reality, all arguments from experience are founded on the similarity which we discover among natural objects, and by which we are induced to expect effects similar to those which we have found to follow from such objects. ... From causes which appear *similar* we expect similar effects. This is the sum of all our experimental conclusions."

Thus Hume suggests the notion of similarity as key to the procedure by which we use cases. While Hume focuses on causes and effects, we also highlight the similarity between acts. In the most general formulation, our similarity function is defined over cases. However, if we wish to restrict similarity judgments to those objects that are known at the time of decision, the similarity function should be defined over pairs of problems, or, at most, problem-act pairs.

It is important to note that the similarity function is not assumed to be "known" by the decision maker in the same sense cases are. While cases are claimed to be "objectively known," the similarity is a matter of subjective judgment. Where does it come from, then? Or, to be precise, what determines our similarity judgments (from a descriptive viewpoint), and what should determine it (taking a normative view)?

Unfortunately, we cannot offer any general answers to these questions. However, it is clear that the questions will be better defined, and the answers easier to obtain, if we know how the similarity is used in the decision making process. Differently put, the similarity function, being a theoretical construct, will gain its meaning only through the procedure in which it is used.

One relatively obvious candidate for a decision procedure is a "nearest neighbor" approach. It suggests that, confronted with a decision problem, we should look for the most similar problem which has been encountered in the past. Unfortunately, this approach does not seem to be very satisfactory: if one is happy with the result of the "nearest" case, one may indeed choose the same act that was chosen in that case. But which act is to be chosen if the "nearest" case ended up yielding a disastrous outcome? Furthermore, assume that the nearest case resulted in a reasonable outcome, but in many other cases, that are somewhat less similar to the problem at hand, a different act was chosen, and it yielded particularly good outcomes in all of them. Is it still reasonable, from either descriptive or normative point of view, to choose the act that was chosen in the *most* similar case?

It therefore appears that a more sensible procedure would not rely solely on the similarity function, nor would it depend only on the most similar case. First, one needs to have a notion of *utility*, i.e., a measure of desirability of outcomes, and to take into account both the similarity of the problem and the utility of the result when using a past case in decision making. Second, one should probably use as many cases as deemed relevant to the decision at hand, rather than the "nearest neighbor" alone.

The CBDT functionals now appear as natural candidates to incorporate both similarity and utility; they prescribe that acts be evaluated by a similarity-weighted sum of the utility they yielded in past cases. However, it should not come as a great surprise to us if in certain circumstances the additive separability of the formula above will prove to be too restrictive. (In particular, see the discussion in Section 6 below.)

The functionals described in Section 2 evaluate acts by resorting to the past performance of acts. The reader might wonder, how do cases that do not involve acts enter the CBDT functionals? Indeed, these functionals do not seem to capture very sophisticated reasoning. In Section 5 we extend our theory to deal with planning. In the extended model, cases that simply relate causes to effects, without the decision maker's intervention, will play a role in her planning, to the extent that they affect the evaluation of future acts.

4.3 Do CBDM's Know Rules?

Equipped with knowledge of past cases, similarity judgments and utility valuations, case-based decision makers (CBDM's) may go about their business, taking decisions as the need arises, without being bothered by induction and general rules, and without ever having to deal with inconsistencies. Yet, if they are asked why they made a certain decision, they are likely to give answers in the form of rules. For instance, if you turn a door's handle in order to open it, and you are asked why you chose to turn the handle, it is unlikely that your answer would be a list of particular cases in which this trick happened to work. You are most likely to say, "Because the door opens when one turns the handle," that is, to formulate your knowledge as a general rule.

In a sense, then, case-based decision makers often behave *as if* they knew certain rules. Furthermore, whenever the two descriptions are equivalent, the language of "rules" is much more efficient and parsimonious than that of "cases." However, the advantage of a case-based description is in its flexibility. In the same example given above, suppose that a person (or, say, a robot we are now programming) finds out that the door refuses to open despite the turning of the handle. A rule-based decision maker, while struggling with the door, also undergoes internal, mental commotion. Not only is the door still shut, one's rule base has been contradicted. Some rules will have to be retracted, or suspended, before our decision maker will be able to use the second rule "If turning the handle doesn't work, call the janitor." By contrast, a CBDM has only the door to struggle with. The failure of the turn-the-handle act will make it less attractive, and will make another act the new U-maximizer. Perhaps the very failure of the first act will make the whole situation look more similar to other cases, in which only a janitor could open the door. At any rate, a CBDM does not have to apply any special procedures to re-organize her knowledge base. Like the rule-based one, the case-based decision maker also has to wait for the janitor; but she does so with peace of mind.

In other words, the rules-vs.-cases choice faces us with a familiar tradeoff between parsimony and accuracy. Rules are simply described, and they are therefore rather efficient as a means to represent knowledge; but they tend to fail, namely to be contradicted by evidence and by their fellow rules. Cases tend to be numerous and repetitive, but they are never a source of inconsistency. In view of the theoretical problems associated with rules, it appears that case-based models are a viable alternative.

4.4 Two Roles of Rules

While CBDT rejects the notion of "rules" as objects of knowledge, it may still find them a useful tool. Even if one is convinced that they are too crude to be "correct," rules may still be convenient approximations of cases. Furthermore, they provide a language for efficient information transmission. As such, rules can have two roles:

(i) A rule may summarize many cases. If we think of a rule as an "ossified case," (Riesbeck and Schank (1989)) it is natural to imagine one individual (system) telling another about many cases by conveying a single rule that applies in all of them;

(ii) A rule may point to similarity among cases. That is, even if two people (systems) have the same cases in their memory, one may be unaware of certain common denominators among them. Especially when the amount of information is vast, an abstract rule may help in finding analogies. For instance, claims such as "Peace-keeping forces can succeed only if the belligerent forces want them to" or "The stock market always plunges after presidential elections" serve mainly to draw the reader's attention to known cases, rather than to tell her about unknown ones. Furthermore, many "laws" in the social sciences, though formulated as rules, should be thought of as "proverbs:" they do not purport to be literally true. Rather, their main goal is to affect similarity judgments. In this capacity, the fact that rules tend to contradict each other poses no theoretical difficulty. Indeed, it is well-known that proverbs are often contradictory.[3] Once they are incorporated into the similarity function, the latter will determine which prevails in each given decision situation.

To sum, CBDT may incorporate rules, and experts' knowledge formulated as rules, either as a summary of cases or as a similarity-defining feature of cases. Yet, within the framework of CBDT, rules are not taken literally, they are not assumed "known," and their contradictions are blithely ignored.

5. Planning

CBDT describes a decision as a single act that directly leads to an outcome. In many cases of interest, however, one may take an act not for its immediate outcome, but in order to be in a position to take another act following it. In other words, one may plan ahead. In this section we extend CBDT to a theory of case-based planning.

[3] The notion of a "rule" as a "proverb" also appears in Riesbeck and Schank (1989). They distinguish among "stories," "paradigmatic cases," and "ossified cases," where the latter "look a lot like rules, because they have been abstracted from cases." Thus, CBR systems would also have "rules" of sorts, or "proverbs," which may, indeed lead to contradictions.

The formal model of CBDT distinguishes between problems, acts, and results. When planning is considered, the distinction between problems and results is blurred. The outcome of today's acts will determine the decision problem faced tomorrow. Thus the formal model of case-based planning will not distinguish between the two. Rather, we employ a unified concept of a "position," that also can be viewed as a set of circumstances. A position might be a starting point for making a decision, that is, a "problem," but also the end "result." We will therefore endow a position with (i) a set of available acts (in its capacity as a "problem") and (ii) a utility valuation (when considered a "result"). Part of the planning process will be the decision, whether a certain position should be a completion of the plan, or a starting point for additional acts (or a sub-plan).

Let P be a finite and non-empty set of *positions*. We assume that it is endowed with a strict partial order $\succ \subset P \times P$, interpreted as "is later than." Let A denote a finite and non-empty set of *acts*. For $p \in P$, let $A_p \subset A$ be the set of acts *available* at p. We introduce $a_0 \in A$ to be interpreted as "do nothing," that is, as the "null act," and assume that $a_0 \in A_p$ for all $p \in P$. The set of *cases* is

$$C \equiv \left\{ (p,a,q) \in P \times A \times P \,\middle|\, a \in A_p, q \succ p \right\}.$$

A decision maker is characterized by a *utility* function $u : P \to \Re$ and a *similarity* function $s : C \times C \to \Re_+$. A position p_0 can be viewed as posing a decision problem for which it is the "initial position." A plan is an assignment of acts to a subset of positions that includes the initial position. Formally, a *plan* for a position (or "problem") p_0 is a pair (N,δ) where $N \subset P$ and $\delta : N \to A$ satisfy: (i) $p_0 \in N$; and (ii) $\delta(p) \in A_p$ for all $p \in N$. It will prove convenient to extend δ to all positions in P by setting $\delta(p) = a_0$ for all $p \notin N$. (Alternatively, one may define δ over P to begin with, and set N to be the positions for which δ does not assign the null act a_0. However, we find the present formulation more intuitive, since it highlights the subset of positions that are salient to the decision maker.)

Given a plan (N,δ) for a position p_0, we are interested in its evaluation based on *memory*, $M \subset C$. First define, for any case $c \in C$, the support that M lends to c by

$$S_M(c) = \sum_{c' \in M} s(c,c')\ .$$

We extend S_M, defined on cases, to all triples in $P \times A \times P$ be setting it to zero whenever its argument in not in C.

Next, consider the weighted directed graph $G = G_{(N,\delta,M)} = (P,E,w)$ defined as follows.

$$E = \left\{ (p,q) \in P^2 \mid q \succ p \right\}$$

and

$$w(p,q) = \frac{S_M\big((p,\delta(p),q) \big)}{\sum_{q'} S_M\big((p,\delta(p),q') \big)}$$

for $(p,q) \in E$, provided that the denominator does not vanish, and zero otherwise. We may view the function w as transition probabilities of an acyclical Markov chain, whose "states" are the positions. For this interpretation, we complement its definition by $w(p,p) = 1$ whenever $\sum_{q'} S_M\big((p,\delta(p),q')\big) = 0$.

Let $f = f_{(p_0,N,\delta,M)} : P \to [0,1]$ be defined as follows: for $q \in P$, $f(q)$ is the probability that a process, starting at p_0 and governed by w, will be absorbed in q. The plan (N,δ) is evaluated by

$$V(N,\delta) = \sum_{q \in P \setminus N} f(q)u(q) \quad .$$

In the Appendix we provide an axiomatic derivation of this rule.

Note that the probabilities $f(q)$ in this summation need not add up to one. A position $p \in N$ for which $\sum_{q'} S_M\big((p,\delta(p),q')\big) = 0$ has a weight $f(p)$ which is not taken into account in the above expression. This is in line with the CBDT model, where an act that was not previously encountered is assigned an "aspiration level," and the latter is normalized to be zero. Finally, notice that V depends on memory, on the initial position, as well as on the primitives of the model. Namely, $V = V_{(P,\succ,A,u,s,p_0,M)}$.

The definition of a "plan" does not require that every position, at which an action is planned, be reachable from the initial position p_0 by positive weight arcs in G. It is easy to see that the evaluation of a plan depends only on those positions in it that are indeed reachable from p_0 in this sense. However, the present formulation encompasses "incomplete" plans: a tentative plan may be devised for the contingency of reaching position q, even if it is not yet clear how the latter can be reached from p_0.

The transition probabilities are defined also for positions that are not part of the plan. Thus, if the decision maker has a reason to believe that a position she would like to get to would evolve, by no action of hers, into another position, the evaluation functional forces her to take this further development into account. The planner's ability to foresee the future is captured by the similarity function. As in CBDT, this function incorporates such factors as probability of recall, on top of "intrinsic" similarity judgments. Correspondingly, a decision maker who does not know of certain cases, or who does not bother to think about them, would be modeled as having zero similarity values to the relevant cases. However, should the decision maker be aware of such cases, the decision rule described above does not allow her to ignore them by simply omitting them from the plan's domain N.

In the above formulation, memory is modeled as a set of cases. Alternatively, one may think of it as a sequence of cases, indexed by natural numbers. Such a formulation may complicate notation, but it tends to deal more gracefully with repeated cases. In the present (set) formulation, it is probably simplest to assume that no position can ever be encountered twice. While our definitions above make sense even when repetitions are allowed, one may (implicitly) assume that the description of a "position" is elaborate enough to contain a time parameter, the protagonist's identity, and so forth. In this case no position may appear in memory more than once, and "basic identity" of positions is reflected in the similarity function.

Our notion of a "position" is closely related to the notion of a "state" in the literature on Markov chains and dynamic programming, as well as in the planning literature. Furthermore, out evaluation functional bears resemblance to the evaluation of "strategies" for "decision trees." However, there are a few differences between our theory and Bayesian planning.

Our evaluation of plans differs from the Bayesian approach in the same way V-maximization (in one-stage CBDT) differs from expected utility theory. That is, the weights assigned to transitions in our graph are not standard probabilities. As opposed to the classical notion of "relative frequencies," these weights are derived from *similar* cases. The interesting applications we have in mind do not involve repetitions of the same problems in more-or-less the same conditions, allowing for standard statistical techniques. Rather, we view this application (in which the similarity values are either 0 or 1) as a very special case; we will resort to it as a benchmark, but it is hardly the motivation for developing our theory. On the contrary, CBDT may be considered an attempt to generalize theories of decision making under uncertainty from this special and well-studied instance to the domain in which statistics is of little help.

An extreme example in which statistical data are scarce is the case of almost novel situations, about which practically no past data are available. One of the standard approaches in statistics is to resort to Bayesianism: to follow Bayes' in arguing that one should have (subjective) beliefs over any unknowns. As explained above, we find this view unrealistic. The way CBDT deals with novel situations is to assign a default "aspiration level" value to unknown outcomes. Correspondingly, our evaluation of plans differs from Bayesian planning (or dynamic programming) in that it allows some of the probability mass to "disappear." More precisely, the utility is scaled so that the aspiration level is set to zero, and unknown outcomes are assigned this aspiration level (which is equivalent to ignoring them in the summation).

As opposed to decision-trees analysis, our model does not distinguish (a-priori) between a "decision node" and a "terminal node" or an "outcome." Both are simply "positions." It is the decision maker's plan which introduces this distinction. Thus, one decision maker may have a more elaborate plan than another. Alternatively, the same decision maker may take a tentative plan and further elaborate it. We believe that the model presented above is more cognitively plausible than the model of decision trees, in that it does not require the decision maker to have the complete tree in her mind. Rather, the decision maker imagines only those positions to which her similarity values assigns positive weight. As she thinks about the problem further, or as a result of new information, she might be aware of more cases, have more positive-weight arcs in the planning graph, consider "terminal" positions as starting points for further planning, and so forth.

We believe that our model, especially if taken to represent a dynamic planning process, is a much better descriptive theory of human planning than the decision-tree model. But even from a normative viewpoint we find that it may be advantageous: in complex and new situations, the complete decision tree might be huge, and might require probabilistic evaluations that have little data to be based on. By contrast, our planning model might be a better tool for actual planning. It might even be viewed as a conceptual framework within which one can derive and refine decision trees.

Dynamic programming models also do not distinguish between a "decision node" and an "outcome." They include "states," where the system is at a given state at every time. Payoffs are collected by the decision maker along transition arcs, and are typically assumed to be evaluated by a discounted sum. These models are theoretically very appealing, but they suffer from a number of shortcomings. First, they require knowledge of probabilities. Hence their normative appeal is marred when there is no sound source for the generation of the latter. Needless to say, in those cases their descriptive value is also limited. Second, they are descriptively questionable because they pre-suppose a very high degree of rationality, in imagining all payoffs along the various paths and aggregating them by (infinite) discounted sums. By contrast, our planning model assumes that the perceived payoffs only appear at "terminal" positions, ignoring the paths that led to them. In cases where the path affects payoffs in a significant way, our model may have to treat the terminal position differently, incorporating the path payoff into the terminal position's payoff. As long as this is the exception rather than the rule, we believe that our model is a viable alternative to dynamic programming.

How "rational" is it to choose a plan that maximizes V? One relatively weak criterion of rationality is dynamic consistency: suppose the decision maker has chosen a V-maximizing plan and started following it. Now she finds herself in a new position. She can stick to the original plan, or she can re-optimize and choose a plan that is optimal for the new problem. If it so happens that she has to choose a new plan, her original choice was not dynamically consistent. Indeed, one may wonder, why has she not planned to re-optimize at the outset? As long as no new information is provided, why could she not envisage her choice of a new plan already at the original position?

The following observation states that V-maximization is, indeed, dynamically consistent.

<u>Observation</u>: Let there be given a model $(P,\succ,A,u,s,p_0,M)$ and assume that (N,δ) is an optimal (V-maximizing) plan for p_0. Let $p \in N$ be a position reachable from p_0 using positive similarity paths. Then (N,δ) is also optimal for p.

<u>Proof</u>: Suppose that a different plan, (N',δ') were strictly better than (N,δ) at p. Consider the plan (N'',δ'') that "mimics" (N',δ') at the sub-graph defined by p, and

follows (N,δ) elsewhere. Since p is reachable from p_0, (N'',δ'') is strictly better than (N,δ) at p_0. •

We conclude this section with two simple results regarding computational complexity. Specifically, we show that the following two problems can be solved in polynomial time:

1. V Calculation: Given a model $(P,\succ,A,u,s,p_0,M)$ and a plan (N,δ), find its V value.

2. V Maximization: Given a model $(P,\succ,A,u,s,p_0,M)$, find a plan (N,δ) that is optimal (V-maximizing) for it.

Proposition 1: V Calculation can be solved in polynomial time.

Proof: Given $\succ$, find a linear order $\succ'$ on P that agrees with $\succ$. Following $\succ'$, compute the probability that the Markov process visits any particular position, and thus also the function f. Finally, compute V given f. It is easy to see that each of these steps requires a number of operations that is polynomial in $|P|$. •

Proposition 2: V Maximization can be solved in polynomial time.

Proof: In view of the dynamic consistency result, one may start with an optimal plan for $\succ$-maximal positions, and use a "dynamic programming" (or "backward induction") technique. •

6. CBDT and Induction

Whereas CBDT does not involve explicit induction, namely, the generation of rules from instances thereof, it does engage in learning from the past regarding the future. Thus it can be said to involve implicit induction, or "extrapolation." We devote this section to the process of induction as viewed from a case-based perspective. In the first sub-section we distinguish between two levels of (implicit) induction within the CBDT model. In the second we compare implicit and explicit induction as descriptive theories of the way people extrapolate from past cases.

6.1 Memory-Dependent Similarity and Two Levels of Induction

In the model of Section 2 the similarity function is assumed to be memory-independent. However, the similarity function may also depend on the problems that were encountered in the past, as well as on the results obtained in them.

We start with the following example. Consider a decision maker who has two coffee shops in her neighborhood, 1 and 2. Once in a shop (which determines a "problem"), she has to decide what to order (which act to choose). In the past, she has visited both of them once in the morning (M) and once in the evening (E), ordering "cafe latte" in each of these four problems. The four problems she recalls are: $(M1, M2, E1, E2)$. (Notice that which shop to go to is *not* a decision variable in this story.) Now assume that the quality of the coffee she had was either 1 (high) or −1 (low). Let us compare two possible memories: in the first, the result sequence is $(1,1,-1,-1)$, while in the second it is $(1,-1,1,-1)$. In the first case the decision maker would be tempted to assume that what determines the quality of cafe latte is the time of the day. Correspondingly, she is likely to put more weight on this attribute in future similarity judgments. On the other hand, in the second case, the implicit induction leads to the conclusion that coffee shop 1 simply serves a better latte than 2, and more generally, that the coffee shop is a more important attribute than the time of the day. In both cases, the way similarity will be evaluated in the future depends on memory as a whole, including the results that were obtained.

Generally, one may distinguish between two levels of inductive reasoning in the context of CBDT. First, there is "first order" induction, by which similar past cases are implicitly generalized to bear upon future cases, and, in particular, to affect the decision in a new problem. The version of CBDT presented here attempts to model this process, if only in a rudimentary way. However, there is also "second order" induction, by which the decision maker learns not only what to "expect" in a given case, but also *how to conduct first-order induction*, namely, how to judge similarity of problems. The current version of CBDT does not model this process. Moreover, it implicitly assumes that it does not take place at all.

Specifically, one would expect that when some process of "second-order induction" affects similarity judgments, there would be some plausible counterexamples to U-maximization. Indeed, consider the following example: coffee shops 1, 2, 3, and 4 serve both cafe latte and cappuccino. Shops 1 and 2 are in our decision maker's neighborhood, and she had the opportunity to try both orders in each of them, both in the morning and in the evening. Shops 3 and 4 are in a different town, and probably bear little resemblance to either 1 or 2. So far the decision maker has only tried the latte in 3 in the afternoon (A), and the cappuccino in 4 at night (N). Both resulted in high-quality coffee. The next afternoon she is in shop 4, trying to decide what to order. Based on her experience, both acts are likely to have a positive U-value. Yet, she may still distinguish between them depending on her similarity function. If she puts more weight on the time of the day, the latte, which was successfully tried in the afternoon, is a more promising choice; if, however, she tends to "believe" that similarity is mostly determined by the shop, she should perhaps order cappuccino, as she did yesterday night in the same shop. Let us now consider the following vectors (where empty entries denote zeroes):

act	Problems									
profiles	$M1$	$M2$	$E1$	$E2$	$M1$	$M2$	$E1$	$E2$	$A3$	$N4$
x	1	1					−1	−1	1	
y			−1	−1	1	1				1
z	1		1			−1		−1	1	
w		−1		−1	1		1			1
d		−1	1			−1	1			

In this example memory contains ten cases. It is convenient to think of the ten problems as formally distinct: for instance, each may be uniquely identified by a time parameter. However, in the table we suppress this parameter and specify only the features of the problem that are deemed relevant, namely the coffee shop and the time of the day.

We further assume that in each of the problems only two acts – "cafe latte" and "cappuccino" – were available. The vectors x, y, z, w, and d are "act profiles;" that is, they designate a conceivable history of an act. If the act was not chosen in a particular case, it is assigned a default "utility" value of zero. If it was, it is assigned the actual utility that resulted from it in this case.

We would now like to consider the preferences between cafe latte and cappuccino under two separate scenarios. In the first, the preference question would reduce to comparing x and y, while in the second – to comparing z and w. Suppose first that x is the act profile of "cafe latte" and y – of "cappuccino." Focusing on the first two rows in the table, the results obtained in the first eight problems clearly indicate that it is the time of the day that matters: all morning coffees were high quality, all evening ones were low quality. Based on this "general observation," the decision maker has learnt to appreciate the crucial role of the time of the day, and she is unlikely to put too much weight on the "night problem" $N4$ when making a decision in the afternoon. Thus, she expresses a preference for x over y when faced with the problem $p = A4$.

By a similar token, when comparing z and w, the decision maker concludes that the shop is very important, but the time of the day does not really matter. Hence she puts more weight on the experience in the same shop – problem $N4$ – and decides to order cappuccino, i.e., she prefers w over z (for the same decision problem $p = A4$).

It is easily verified that this preference pattern is inconsistent with U-maximization for a fixed similarity function s. Indeed, for any such function s, since $z - x = w - y = d$, we have

$$U(z) - U(x) = U(w) - U(y) = U'(d)$$

from which we derive

$$U(x) - U(y) = U(z) - U(w).$$

That is, x is preferred to y *if and only if* z is preferred to w, in contradiction to the preference pattern we motivated above. Thus second-order induction may result in violations of CBDT as presented above.

Similar examples may be constructed, in which the violation of U-maximization stems from learning that certain *values* of an attribute are similar, rather than that the attribute itself is of importance. That is, instead of learning that the coffee shop is an important factor, Agent may simply learn that coffee shop 1 is similar to coffee shop 2. Similarly, one may construct examples in which intuitive preferences patterns violate maximization of the other linear CBDT functionals.

One obvious drawback of the functional U that is highlighted here is the fact that it is additively separable across cases. Specifically, second-order induction renders the "weight" of a set of cases a non-additive set function. Since several cases *in conjunction* may implicitly suggest a "rule," the effect of all of them together may exceed the sum of their separate effects. Differently put, the "marginal contribution" of a case to overall preferences depends not only on the case itself, but also on the other cases it is lumped together with. For instance, a utility value of 1 in problem

$M1$ has a different effect when coupled with the value 1 in problem $E1$ (as in vector z) than it has when coupled with the same value in $M2$ (as in x).

A possible generalization of additive functionals that may account for this "non-additivity" involves the use of non-additive measures, where aggregation of utility is done by the Choquet integral. (Choquet (1953-4). See also Schmeidler (1989), who introduced this technique to decision making under uncertainty.) However, it should be noted that when second-order induction takes place, it is not only the case-additivity assumption that is being challenged. With similar examples one may convince oneself that the very assumption that preference between acts is determined solely by their "act profiles" may fail in the presence of inductive learning of the similarity function. For instance, consider a matrix as above, where the acts chosen in the first eight problems were neither "cafe latte" nor "cappuccino," but rather two different ones, that yielded the results given by the table. The decision maker would still draw the same general conclusions about the relative importance of the two attributes of a "problem," and her preference between the latte and the cappuccino would thus depend on all of her memory, including the act profiles of other acts. In particular, second-order inductive reasoning is one plausible example in which case-based preferences do not necessarily satisfy "independence of irrelevant alternatives." That is, the preference between two acts may change when other acts are introduced into the choice set, even if the latter are considered "worse choices" than both of the former.

The distinction between the two levels of induction may be extended to the process of learning and the definition of "expertise." A case-based decision maker learns in two ways: first, by introducing more cases into memory; second, by refining the similarity function based on past experiences. By learning more cases, our decision maker obtains a wider "data base" for future decisions. This process should generally improve her decision making. Of course, the cases learnt may be biased or otherwise misleading; yet, one may expect that, as a rule, and barring computational costs, the knowledge of more cases leads to a "better" first-order induction as embodied in case-based decision making.

This improvement of the first-order induction may be viewed as "quantitative." That is, to the extent that CBDT performs implicit first-order induction, it does so even with a meager memory. Thus the introduction of more cases does not change first-order induction in a fundamental or even qualitative way; it does so only quantitatively. (The term "quantitative" may be misleading, since memories are only partially ordered by inclusion; but the addition of cases has a flavor of "more of the same.") On the other hand, second-order induction may be viewed as a *qualitative* improvement of first-order induction. That is, refining the similarity judgment introduces a new dimension to the process of learning. Rather than simply knowing more, it suggests that a better use may be made of the knowledge of the same set of cases.

Knowledge of cases, which we may dub "type I knowledge," is relatively "objective." Though cases may be construed in different ways, there seems to be relatively little room for dispute about them, since they purport to be "facts." By contrast, knowledge of the similarity function, which we refer to as "type II knowledge," is inherently subjective. Correspondingly, it is easier to compare people's type I knowledge than it is to compare type II knowledge. While even knowledge of type I cannot be easily quantified, it does suggest a clear definition of "knowing more," namely, having a memory that is larger (as defined by set inclusion). On the other hand, it is much more difficult to provide a formal definition of "knowing more" in the sense of "having a better similarity function." It seems that what is meant by that is a similarity function that resembles that of an expert, or one that in hindsight can be shown to have performed better in decision making.

The two roles that rules may play in a case-based knowledge representation system correspond to the two types of knowledge, and to the two levels of induction. Specifically, the first role, namely to summarize many cases, may be thought of as succinctly providing knowledge of type I. Correspondingly, only first-order induction is required to formulate such rules: given a similarity function, one simply lumps similar cases together and generates a "rule."[4] By contrast, the second role – drawing attention to similarity among cases – may be viewed as expressing type II knowledge. Indeed, one needs to engage in second-order induction to formulate these rules: it is required that the similarity be *learnt* in order to be able to observe the regularity the rule should express.

Similarly, "expertise" also has two aspects. First, being an "expert" in any given field typically involves a rich memory, the acquaintance with many cases, or, in short – knowledge of type I. However, an expert can also do more with the same information. That is, (s)he has a more "accurate" and/or more "sophisticated" similarity function, and in our terminology – possesses "more" (or "better") knowledge of type II.

[4] Notice, however, that this is first-order *explicit* induction, i.e., a process that generates a general rule, as opposed to the *implicit* induction performed by CBDT.

These distinctions may also have implications for the implementation of computerized systems. A case-based expert system would typically involve both types of knowledge. The discussion above suggests that it makes sense to distinguish between them. For instance, one would like to separate the "hard," "objective" type I knowledge that may be learnt from an expert from the "soft" and "subjective" type II knowledge provided by the same expert. The first is less likely to change than the second. Furthermore, one may wish to use one expert's knowledge of cases with another expert's similarity judgments.

As a final remark, we would like to draw the reader's attention to the fact that even in the presence of second-order induction, case-based knowledge representation incorporates modifications in a "smooth" way. That is, one may sometimes wish to update the similarity values; this may lead to different decisions based on the same set of cases. But this process does not pose any theoretical difficulties such as those entailed by explicit induction.

6.2 Two Views of Induction: Hume and Wittgenstein

How do people use past cases to extrapolate the outcomes of present circumstances? Wittgenstein (1922, 6.363), for instance, argued that

> "The procedure of induction consists in accepting as true the *simplest* law that can be reconciled with our experiences."

The notion of "simplicity" may be very vague and subjective. (See, for instance, Sober (1975) and Gärdenfors (1990).) Gilboa (1990) suggests employing Kolmogorov's complexity measure for the definition of "simplicity." Using this measure, Wittgenstein's claim may be re-formulated to say that people tend to choose a theory that has a shortest description in a given programming language. We will refer to this theory as "simplicism." Its prediction is well-defined, but no less subjective than the notion of "simplicity." Indeed, it merely translates the choice of a complexity measure to the choice of the "appropriate" programming language.

By contrast, Hume argues that "from similar causes we expect similar effects." That is, that the process of implicit induction, or extrapolation, is based on the notion of similarity, rather than on simplicity. Needless to say, "similarity" is just as subjective as "simplicity," or as "the appropriate language."

If we take simplicism as a formulation of Wittgenstein's view, and CBDT – as a formulation of Hume's, we have a formal basis on which the two may be compared as theories of human extrapolation or prediction. While the following discussion is not unrelated to the comparison of rule-based and case-based methodologies in Section 4, our focus here is on descriptive scientific theories, rather than on knowledge representation technologies.

Two caveats are in order: first, any formal model of an informal claim is bound to commit to a certain interpretation thereof; thus the particular models we discuss may not do justice to the original views. Second, since both "similarity" and "language" are inherently subjective, much freedom is left in the way the two views are brought to bear on a particular example. Yet, we hope that the analysis of a few simple examples might indicate some of the advantages of both views as theories of human thinking.

Consider a simple learning problem. Every item has two observable attributes, A and B. Each attribute might take one of two values, say, A and $\overline{A}$ for A, and B and $\overline{B}$ for B. We are trying to learn a "concept" Σ that is fully determined by the attributes. That is, Σ is a subset of $\{AB, A\overline{B}, \overline{A}B, \overline{A}\overline{B}\}$. Each item poses a "problem" or a "question" (that is one of AB, $A\overline{B}$, $\overline{A}B$, or $\overline{A}\overline{B}$). We are given a few positive and/or negative examples to learn from – that is, items that are either known to be in Σ ("+") or known not to ("–"), and are asked to extrapolate whether the next item is in Σ, based on its observable attributes. At any given time, the set of examples we have observed, or our "memory," may be summarized by a matrix, in which "+" stands for "such items are in Σ," "–" – for "such items are not in Σ," and a blank space is read as "such items have not yet been observed." Finally, a "?" would indicate the next item we are asked about. For instance, the matrix

1	B	$\overline{B}$
A	+	
$\overline{A}$		?

describes a data base in which a positive example AB was observed, and we are asked about $\overline{A}\overline{B}$.

What should we guess $\overline{A}\overline{B}$ to be? Not having observed any negative example, the simplest theory in any reasonable language is likely to be "All items are in Σ," predicting "+" for $\overline{A}\overline{B}$. Correspondingly, if we assume that all items bear some resemblance to each other, a case-based extrapolator will also come up with this prediction.

Next consider the matrices

2	B	$\bar{B}$
A	+	−
$\bar{A}$		?

3	B	$\bar{B}$
A	−	+
$\bar{A}$		?

In matrix 2, the simplest theory would probably be "If B then in Σ, else – not in Σ," predicting that $\bar{A}\bar{B}$ is not in Σ. The same prediction would be generated by CBDT: since $\bar{A}\bar{B}$ is more similar to $A\bar{B}$ than to AB, the former (negative) example would outweigh the latter (positive) one. Similarly, simplicism and CBDT will concur in their prediction for matrix 3. Assuming that attributes A and B are symmetric, we will get similar predictions for the following two matrices:

4	B	$\bar{B}$
A	+	
$\bar{A}$	−	?

5	B	$\bar{B}$
A	−	
$\bar{A}$	+	?

However, the two methods of extrapolation might also be differentiated. Consider the matrices

6	B	$\bar{B}$
A	+	?
$\bar{A}$		−

7	B	$\bar{B}$
A	+	
$\bar{A}$	?	−

The observations in both matrices are identical. The "simplest rule" that accounts for them is not uniquely defined: the theory "If A then in Σ, else – not in Σ," as well as the theory "If B then in Σ, else – not in Σ" are both consistent with evidence, and both would be minimizers of Kolmogorov's complexity measure in a language that has A and B as primitives. (As opposed to, say, their conjunction.) Moreover, each of these simplest theories would predict a positive example in one matrix and a negative one in the other. By contrast, a similarity-weighted aggregation of past examples would leave us undecided between a positive and a negative answer in both matrices.

The CBDT answer, namely, being indifferent between making a positive prediction and making a negative prediction in matrices 6 and 7, appears more satisfactory than the simplest-theory answer. Indeed, in both matrices the evidence for and against a positive prediction are precisely balanced. In a way that parallels our discussion in Section 4, we find that CBDT behaves more "smoothly" at the transition between different rules. Since CBDT uses quantitative similarity judgments, and produces quantitative evaluation functionals, it deals with indifference more graciously than "simplest theories" or "rules."

In these examples it is very natural to suggest that simplicism be interpreted to mean some random choice among competing theories, or an "expected prediction" (of a theory chosen randomly). Indeed, in matrices 6 and 7 above, if we were to take an average of the predictions of the two simplest theories, we will also be indifferent between a positive and a negative prediction. However, if we allow weighted aggregation of theories, we would probably not want to restrict it to cases of absolute indifference. For instance, if ten theories with (Kolmogorov) complexity of 1,001 (say, bits) all agree on a prediction, but disagree with the unique simplest theory, whose complexity is 1,000, it would be natural to extend the aggregated-prediction method to this case as well, despite the uniqueness of the "simplest" theory. But then we are led down the slippery path to a Bayesian prior over all theories, which we find cognitively implausible.

A starker example of disagreement between CBDT and simplicism is provided by the following matrix.

8	B	$\overline{B}$
A	+	+
$\overline{A}$	–	?

The simplest theory that accounts for the data is "A iff in Σ," predicting a negative answer for $\overline{A}\overline{B}$. What will be the CBDT prediction? Had AB not been observed, the situation would have been symmetric to matrices 6 and 7, leaving a CBDT-predictor indifferent between a positive and a negative answer. However, the similarity between AB and $\overline{A}\overline{B}$ is positive, as argued above. (If it is not, CBDT would have predicted "not in Σ" in matrix 1.) Hence the additional observation tilts the balance in favor of a positive answer.

While the CBDT prediction in matrix 8 is hardly intuitive, it is not entirely clear that this example is relevant to the comparison of the views of Hume's and of Wittgenstein's. The CBDT prediction in matrix 8 was "computed" based on the CBDT predictions in the other examples, using the additivity of the CBDT functionals. In other words, matrix 8 is not necessarily a counter-example to Hume's dictum. It may well be a counter-example to the additive separability we assume in CBDT. Indeed, should one consider more general functionals, the effect of the observation that AB is in Σ might be different when it is added to the other two observations in matrix 8, than when it is the only observation. Moreover, one might argue that the similarity function depends on memory as a whole, and not merely on the compared cases. As in sub-section 6.1, the process of learning may entail learning the similarity function itself.

Note that simplicism also allows "second-order induction:" while a case-based decision maker learns the similarity function, a simplicistic extrapolator might learn the language in which the theories should be formulated. For example, adults tend to put less emphasis than do children on color as a defining feature of a car's quality. This might be modeled as second-order induction in both models: in CBDT, it would imply that the weight attached to color in similarity judgments is reduced; in simplicism, it would be captured by including in the language other predicates, and perhaps dispensing with "color" altogether. In this respect, too, the quantitative nature of CBDT may provide more flexibility than the qualitative choice of language in simplicism.

To sum, the CBDT model appears to be more flexible than simplest-theory or rule-based paradigms. However, there is little doubt that the linearity assumption is too restrictive. It remains a challenge to find a formal model that will capture Hume's intuition and allow quantitative aggregation of cases, without excluding second-order induction and refinements of the similarity function.

7. Concluding Remarks

7.1 We do not purport to provide any general insights into the question of "Similarity – Whence?". From a descriptive point of view, this problem is studied in the psychological literature. (See Tversky (1977), Gick and Holyoak (1980), and Gick and Holyoak (1983), among others.) Taking a normative approach, answers are sometimes given in specific domains in which "cases" are an essential teaching technique (such as law, medicine, and business).

At any rate, we believe that the language of CBDT may also be helpful in dealing with the similarity problem. In particular, second-order induction – as defined in the context of CBDT – may provide some hints regarding the evolution of similarity judgments.

7.2 In the discussion of knowledge representation, our main focus is on that "knowledge" used by people (or machines) in everyday situations. However, one may ask to what extent the case-based model applies to the representation of scientific knowledge, or even mathematical knowledge.

Starting with the latter, we agree with Riesbeck and Schank (1989) that a mathematician's knowledge and reasoning technique is most accurately represented by case-based reasoning. Ideas and solutions are considered "creative" precisely when they cannot be generated algorithmically, that is, when the knowledge base they rely on does not contain a wide enough array of obviously-similar cases to induct rules from. Indeed, an idea is "creative" if it does not resemble any known case, or if it relies on original analogies. In other words, creative thinking requires originality either in the (hypothetical) cases considered (corresponding to type I knowledge) or in the similarity function (i.e., type II knowledge).

On the other hand, the *product* of a mathematician's work is almost by definition in the form of rules. Mathematicians are, by and large, interested in producing theorems, or, at most, counterexamples to conjectured ones. Thus the "mathematical knowledge," the accumulation of which is, supposedly, the "aim" of mathematicians, is closer to rules than it is to cases.

When it comes to the sciences, an expert's knowledge is probably best described, as in mathematics, by cases. The product of the scientific work, it would seem, is generally expected to be in the form of rules. However, the degree to which this partly-implicit goal is achieved varies. Generally it appears that in a simple enough environment, that allows for many almost-identical cases to be observed, rules are indeed formulated. But when the environment tends to be unique, scientific knowledge may also take the form of a collection of cases.

CBDT being a scientific theory, one can hardly fail to ask how it applies to itself. Indeed, it does attempt to be a general, rule-style theory, describing decision making and knowledge representation at large. Thus, to the extent that it fails, its failure should be taken as proof of its virtue.[5]

7.3 While our main interest is in the theoretical aspects of CBDT, we would like to note that case-based models need not be impractical. Indeed, rules appear to be much more efficient than the cases from which they were originally derived. Yet one need not actually program each and every case into memory. For instance, repeated cases may be represented by higher similarity values, thus saving both memory and computation time.

[5] On the other hand, if it happens to be a valid description of reality, it is not refuted; rules which happen to be "true" can be translated back to the collection of cases from which they were derived in the first place.

Appendix: An Axiomatic Derivation of Plan Evaluation Rule

This appendix is devoted to an axiomatic derivation of the plan evaluation procedure described in Section 5. We restrict our attention to numerical evaluation procedures. That is, a procedure for the evaluation of plans is assumed to attach a numerical index to each plan, such that higher indices are associated with "better," or "preferable" plans. We assume that a "procedure" is defined for *all* models as in section 5. That is, a procedure is a function

$$V = V_{(P,\succ,A,u,s,p_0,M)} : \mathrm{N}_{(P,\succ,A,u,s,p_0,M)} \to \Re$$

where $\mathrm{N}_{(P,\succ,A,u,s,p_0,M)}$ is the set of all plans (N,δ) for the decision problem p_0 given the primitives of the model $(P,\succ,A,u,s)$ and memory M.

Let V be such a function. We now turn to state some axioms on V that will be shown to characterize the function defined in section 5 above.

<u>A1 Coincidence with CBDT</u>: For a plan (N,δ) with $N=\{p_0\}$ and $\delta(p_0)=a$, and $w(p,q)=0$ whenever $p \neq p_0$,

$$V(N,\delta)=\sum_{p'\in P,\, p'\succ p_0} S'_M((p_0,a,p'))u(p')$$

where

$$S'_M((p_0,a,p'))=\frac{\sum_{(q,b,q')\in M} s((p_0,a,p'),(q,b,q'))}{\sum_{p''}\sum_{(q,b,q')\in M} s((p_0,a,p''),(q,b,q'))}.$$

A1 states the following: consider a "degenerate" plan that consists of a single act. Assume that no immediate consequence of this act leads to further developments (according to the decision maker's memory). Then the plan should be evaluated as its act would be by the functional V'' in CBDT.

The next two axioms basically state that a plan's evaluation does not depend on the exact description of a "position" or of an "act." We first illustrate and motivate them by examples, and then provide the formal statements.

Consider an agent who wants to buy a new car. In order to have enough money, she plans to sell her old car first. Suppose that she intends to take the act "place an ad in the newspaper" (a). This act may lead to (at least) two positions: one (say, p_1) is "the car is sold to a buyer who lives in the city," while the other (p_2) is "the car is sold to a buyer who lives in a suburb." In both positions, the agent intends to take the same act, namely, "use the money to buy a new car" (b). Moreover, the agent expects that, as long as the selling price is identical in both positions, the success of her plan to purchase a new car will not depend on the way she obtained the money. Assume that this implicit belief is supported by the agent's memory. That is, the act b appears to result in the same positions when taken at p_1 and at p_2. (The precise definition of this condition is given below.)

Suppose now that the modeler (and perhaps the agent herself) replaces the two positions, p_1 and p_2, by a single one, p, which is described as "the car is sold." (We implicitly assume here that any potential buyer lives either in the city or in a suburb.) Since the agent anyway plans to take the same action, it appears that the new plan, in which p_1 and p_2 are "collapsed" into p, should be equivalent to the old one.

In a formal definition of "collapsing," the weight of an arc leading to p should be taken to be the sum of the corresponding arcs leading to p_1 and to p_2 in the original graph. To see why this is the "natural" definition, consider first a situation where city dwellers and suburban are very different creatures (according to our agent's subjective similarity judgment). This means that all cases in memory are either similar to a case leading to p_1 or to one leading to p_2, but not to both. Thus the weights of the arcs leading to p_1 and to p_2 are (normalized) sums of similarity values over disjoint sets of cases. Correspondingly, when the agent discards the distinction between the two positions, the new, "unified" position inherits the similarity weights of both its parents.

A similar reasoning applies when a case in memory may be similar to more than one case in the plan. However, a warning is in order here: while we use the term "similarity" or the function s, it may also be interpreted as "support," "weight of evidence," or "relevance," and we implicitly assume that the judgment of these is also reflected in s. The following example illustrates. Assume that our agent in the example above has but one case in memory, in which a car was sold to a city dweller. This case is, of course, similar to (p_0,a,p_1), that is, to the agent selling her car to a city dweller as well. But in general it may also be similar to (p_0,a,p_2), namely, to the agent selling the car to a suburban. Indeed, it is essential to allow our planners to have some similarity to new cases, if we hope that they would ever exhibit some creativity.

It stands to reason that the similarity of the case in memory to (p_0,a,p_1) exceeds that to (p_0,a,p_2). However, when we consider the similarity to (p_0,a,p)

(where p stands for the unified position), it should be the sum of the first two and thus exceed both. Thus, if we have more details in the description of a position, the "similarity" of the case to a case in memory may be *lower* than with less details, even if these details are identical to those specified in the recalled case. "Pure" similarity judgments generally do not follow this pattern; on the contrary, richness of identical details typically enhances the perceived similarity. Yet a "weight of evidence" function is likely to decrease with specificity, and it is this role that the "similarity" function plays in our model.

We have so far described the operation of "unifying" two positions, which we will dub "parallel collapse." We will require that the plan evaluation function be invariant with respect to this operation. Before turning to the formal definition, we describe the second operation, to be referred to as "serial collapse."

Consider our agent again. She may consider the act a, namely, "place an ad in the newspaper," in more detail, and realize that she first has to find out the newspaper's phone number. That is, she plans to first take act a_1, "find the phone number," which she hopes will bring her to a position p_2, "I know the number," and only then can she choose the act a_2, "call the newspaper and place an ad." Conversely, one may view the act a as a "unification" of the two acts, a_1 and a_2. As opposed to parallel collapse, which unifies two positions but does not alter the acts involved, this operation – "serial collapse" – skips a position, and unifies two successive acts. To be precise, the first act is re-interpreted as a contingent plan, to take a_1 first, and then a_2 if p_2 is reached.

We define the similarity (or "weight of evidence") associated with the arcs involving the new act to be the product of the normalized similarity weights leading to and from the intermediate position p_2. The intuition underlying this definition is best explained in the case where all similarity values (between cases) are 0 or 1. In this case, the normalized sums of similarity values are simply relative frequencies of results (of a given act at a given position) in the past. For relative frequencies (which are defined per position), the product rule is almost tautological. To be precise, if the graph involved is a tree, then this rule follows from the definition of "conditional frequency," just as Bayes' definition of conditional probability implies $P(A \cap B) = P(A)P(B|A)$. In general, however, the product rule implicitly assumes that the distribution of outcomes that may result from taking an act at a given position, does not depend on the path that led to this position. Indeed, should memory indicate that this is not the case, the position should be split into several positions.

As in the case of parallel collapse, we extend this intuition to the more general case, in which the similarity values need not be 0 or 1, and which allows the planner to use cases that are not "identical" to the current problem.

We finally turn to define parallel and serial collapses formally. Let there be given a model $(P,\succ,A,u,s,p_0,M)$ and a plan (N,δ) for it. The verbal description of the collapse operations involves changing the primitives of the model, i.e., re-defining the set of positions and/or of acts. For simplicity of notation, however, we will keep these fixed, and reflect all relevant changes in the similarity function. Specifically, we fix $(P,\succ,A,u,p_0,M)$ and consider evaluation functions $V_s(N,\delta)$, that is, we focus on the way in which the evaluation depends on s.

A few additional definitions will prove useful. Two positions $p_1,p_2 \in P$ are said to be $\succ$*-equivalent* if for all $q \in P$, $q \prec p_1$ iff $q \prec p_2$ and $q \succ p_1$ iff $q \succ p_2$. A position $p \neq p_0$ is *inessential* with respect to the similarity function s if for all $q \in P$ and $a \in A$, $S_M((p,a,q)) = S_M((q,a,p)) = 0$. It is essential otherwise, i.e., if there is a non-zero similarity arc leading to or from it.

We now can state the conditions under which two positions may be collapsed. Assume that $p_1,p_2 \in P$ satisfy:

(i) $p_i \neq p_0$ $(i = 1,2)$;

(ii) p_1 and p_2 are $\succ$-equivalent;

(iii) $\delta(p_1) = \delta(p_2)$;

(iv) For all $q \in P$, $w(p_1,q) = w(p_2,q)$.

The similarity function s' describes a *parallel collapse* of p_2 onto p_1 relative to the similarity function s if, denoting by w' the normalized similarity weights generated by s' and M:

$$w'(q,p_1) = w(q,p_1) + w(q,p_2)$$

$$w'(q,p_2) = 0 \qquad\qquad \text{for all } q \in N;$$

and w' equals w elsewhere.

<u>A2 Parallel Collapse</u>: If, for some p_1, $p_2 \in P$, s' describes a parallel collapse of p_2 onto p_1 relative to the similarity function s, then

$$V_{s'}(N,\delta) = V_s(N,\delta) .$$

Observe that condition (iv) requires that the normalized similarity weights leading out of the two positions be identical. We argue that it captures the intuitive notion of two positions being "practically the same" in the planner's eyes. Consider the car selling example again. While it appears irrefutable that the source of the money should not affect its potential use, in the final analysis this is a conclusion one may only draw from experience. Indeed, it is conceivable that some aggressive creditors of the buyer will show up and ask for the money with which the car was bought; or that our agent will find herself boycotted by city dwellers for having traded with suburbans. In short, our intuition that the seller's identity does not matter is based on our experience. That is, on similar past cases that predict the same results for the act a at p_1 and at p_2.

We now turn to define "serial collapse." Assume that $p_1, p_2 \in P$ and $a_1, a_2 \in A$ satisfy the following conditions:

(i) $p_2 \succ p_1$;
(ii) $\delta(p_i) = a_i$ $(i = 1,2)$;
(iii) for $q \neq p_1$, $w(q, p_2) = 0$.

The similarity function s' describes a *serial collapse* of a_2 at p_2 onto a_1 at p_1 relative to the similarity function s if (using the same notation):

$$w'(p_1, q) = w(p_1, q) + w(p_1, p_2)w(p_2, q) \qquad \text{for all } q \succ p_2;$$

$$w'(p_1, p_2) = 0;$$

and w' equals w elsewhere.

<u>A3 Serial Collapse</u>: If, for some $a_1, a_2 \in A$, $p_1, p_2 \in N$, s' describes a serial collapse of a_2 at p_2 onto a_1 at p_1 relative to the similarity function s, then

$$V_{s'}(N,\delta) = V_s(N,\delta) .$$

Observe that the serial collapse is allowed only if the position that is practically deleted from the graph, namely p_2, is reachable only from p_1 (condition (iii)). Indeed, should p_2 be reachable from other positions, collapsing a_2 at p_2 should be reflected in all position-act pairs leading to it. Note also that the product of the normalized similarity weight of (i) the arc entering p_2 and (ii) that of the arc leading from p_2 to some position q – is *added* to the weight of the arc leading from p_1 directly to q. This is in line with the definition of S_M as a sum over similarities. Indeed, were multiple arcs allowed in our model, we could simply state that a new arc from p_1 to q has a weight that is the product of the weights of its parents, and the summation would be taken care of in the function S_M.

Finally, we need a structural assumption, stating that the set of positions is rich enough. It is a technical assumption, whose role will become clear in the proof. At this point we only mention that we need this assumption because we chose to hold the primitives of the model fixed, and to reflect all "collapses" in the similarity function.

Structural Assumption: Let k be the number of essential positions in P. Then, for each $p \in P$, P contains at least 2^k inessential positions that are $\succ$-equivalent to p.

We can now state:

Proposition: Under the structural assumption, the function $V = V_{(P,\succ,A,a,s,p_0,M)}$ defined above is the unique evaluation function satisfying A1-A3.

Proof: The fact that V satisfies A1-A3 follows from standard probability calculus. We prove that it is uniquely defined by these axioms. Let there be given some function U satisfying A1-A3. We will show that $U = V$ in three steps.

Step 1: Consider first the set of "degenerate" plans defined by one act. (I.e., $N = \{p_0\}$ and $w(p,q) = 0$ whenever $p \neq p_0$.) In this case the graph contains only paths of length 1, and A1 defines U uniquely.

Step 2: Next consider the pairs of plans and similarity functions for which the graph G is a tree. By successive serial collapses, we can find a graph as in step 1, that is equivalent to the original one according to both U and V, by virtue of A3. Since the two functions coincide on all such "degenerate" graphs, they also have to coincide on all trees.

Step 3: Given a general graph, reduce it to a tree (that will be both U- and V-equivalent to it) as follows. For each path leading from p_0 to some $q \in P$, consider an inessential position that is $\succ$-equivalent to q. (The existence of these is guaranteed by the structural assumption.) Any position $q \in P$ can be "split" into several positions, each corresponding to a different path leading from p_0 to q. The "splitting" is done in a way that mirrors parallel collapse. That is, if q' is one of the "new" (previously inessential) versions of q, $w(q',p)$ is set equal to $w(q,p)$ for all p. The resulting graph is a tree, and, by successive applications of A2, has the same U- and V-values as the original one. Thus U and V have to coincide on all graphs. •

Our analysis assumes that plans are mapped to numbers. Presumably, these numbers only matter to the extent that they help rank plans. One may therefore ask, whether the evaluation procedure axiomatized here can be derived from more primitive data, namely, from a preference relation over plans given various memories. While we conjecture that the answer is in the affirmative, we also suspect that such a derivation is bound to be more cumbersome.

REFERENCES

Choquet, G. (1953-4), "Theory of Capacities," Annales de l'Institute Fourier, 5, 131-295.

Gärdenfors, P. (1990), "Induction, Conceptual Spaces and AI," Philosophy of Science, 57, 78-95.

Gick, M. L. and K. J. Holyoak (1980), "Analogical Problem Solving," Cognitive Psychology 12, 306-355.

Gick, M. L. and K. J. Holyoak (1983), "Schema Induction and Analogical Transfer," Cognitive Psychology 15, 1-38.

Gilboa, I. (1993), "Philosophical Applications of Kolmogorov's Complexity Measure," in Philosophy of Science in Uppsala.

Gilboa, I. and D. Schmeidler (1992), "Case-Based Decision Theory," forthcoming in The Quarterly Journal of Economics.

Gilboa, I. and D. Schmeidler (1993), "Case-Based Optimization," forthcoming in Games and Economic Behavior.

Gilboa, I. and D. Schmeidler (1994), "Act Similarity in Case-Based Decision Theory," mimeo.

Hanson, N. R. (1958), Patterns of Discovery. Cambridge, England, Cambridge University Press.

Hume, D. (1748), Enquiry into the Human Understanding. Oxford, Clarendon Press.

Kolodner, J. L., Ed. (1988), Proceedings of the First Case-Based Reasoning Workshop. Los Altos, CA, Morgan Kaufmann Publishers.

Kolodner, J. L. and C. K. Riesbeck (1986), Experience, Memory and Reasoning. Hillsdale, NJ, Lawrence Erlbaum Associates.

Levi, I. (1980), The Enterprise of Knowledge. Cambridge, MA, MIT Press.

March, J. G. and H. A. Simon (1958), Organizations. New York, Wiley.

McDermott, D. and J. Doyle (1980), "Non-Monotonic Logic I," Artificial Intelligence 25, 41-72.

Moser, P. K., Ed. (1986), Empirical Knowledge. Rowman and Littlefield Publishers.

Quine, W. V. (1953), "Two Dogmas of Empiricism," in From a Logical Point of View. Cambridge, MA, Harvard University Press.

Quine, W. V. (1969), "Epistemology Naturalized," in Ontological Relativity and Other Essays. New-York, Columbia University Press.

Reiter, R. (1980), "A Logic for Default Reasoning," Artificial Intelligence 13, 81-132.

Riesbeck, C. K. and R. C. Schank (1989), Inside Case-Based Reasoning. Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.

Savage, L. J. (1954), The Foundations of Statistics. New York, John Wiley and Sons.

Schank, R. C. (1986), Explanation Patterns: Understanding Mechanically and Creatively. Hillsdale, NJ, Lawrence Erlbaum Associates.

Schmeidler, D. (1989), "Subjective Probability and Expected Utility without Additivity," Econometrica, 57, 571-587.

Simon, H. A. (1957), Models of Man. New York, John Wiley and Sons.

Sober, E. (1975), Simplicity. Oxford, Clarendon Press.

Tversky, A. (1977), "Features of Similarity," Psychological Review 84, 327-352.

Wittgenstein, L. (1922), Tractatus Logico Philosophicus. London, Routledge and Kegan Paul.