Report Number: WUCS-93-27

1993-01-01

Human and Machine Cognition Workshop Papers 1989, 1991, 1993

R. P. Loui

Human and Machine Cognition Workshop Papers 1989, 1991, 1993

R. P. Loui

WUCS-93-27

June 1993

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
St. Louis MO 63130-4899

BACK TO THE SCENE OF THE CRIME: OR, WHO SURVIVED YALE SHOOTING?

(1989)

HOW A FORMAL THEORY OF RATIONALITY CAN BE NORMATIVE: IMPLEMENTATION VERSUS INTERPRETATION

(1991)

SHOULD THOSE WHO EXERCISE THE AUTHORITY OF RULES ALSO KNOW THE CASES?

(1993)

BACK TO THE SCENE OF THE CRIME:
OR
WHO SURVIVED YALE SHOOTING?

Invited Paper, read at the International Workshop on Human and Machine Cognition: The Frame Problem, Pensacola, Florida, May 11, 1989.

I

Back to the scene of which crime? Not the shooting of Fred, no, but to the Pylyshyn collection (Z. Pylyshyn. *The Robot's Dilemma*, Ablex, 1986; subsequent quotations are from this text.) of articles on the frame problem: the one in which Hayes says that Fodor "doesn't know the frame problem from a bunch of bananas"; Fodor says that the AI legions are overpaid, or "making fools of themselves working separately", or both; and Hayes rejoins that if the frame problem should not be "left to the hackers, then to whom should it be left: to the philosophers who have not made any progress on it in 2000 years?" Surely this was a crime of at least the proportions of shooting poor Fred.

In Pylyshyn's volume, each author offers a novel diagnosis of the frame problem, as a symptom of a larger representational or methodological problem. Within weeks, Hanks and McDermott introduced the Yale Shooting problem (S. Hanks and D. McDermott. "Default reasoning, non-monotonic logics, and the frame problem," *Proc. AAAI*, 1986.), the unexpected interweaving of *two* frame problems. Who survived Yale shooting? Whose diagnosis was correct? Whose treatment was outdated within weeks of appearing in print?

My working hypothesis had been that the AI contributors, in their eagerness to confine the definition of the problem, missed general principles that the frame problem was to share with Yale shooting. Meanwhile the philosophers suffered blow upon blow from Hayes and McDermott, the keepers of the etymology, but actually survived Yale shooting because of their more distant view.

My conclusion is different. Who survived Yale shooting depends on how charitably we interpret each of these authors. It depends on how broadly we define death. Among the healthy appear to be Hayes and Haugeland. The clear fatality is McDermott. Dennett and Glymour appear to have succumbed to other causes, so the question

of surviving shooting is moot. Janlert, Fodor, and the Dreyfuses are in various states of perforation.

First, to interpret the carnage, it matters what one takes to be the Yale shooting problem, and why essays on the frame problem should be judged in light of Yale shooting.

The Yale shooting problem is sometimes called an instance of the prediction problem. A gun is loaded. We wait. It's fired. Does Fred die because, by default, the gun persisted in being loaded during the wait? Or does Fred persist in being alive, by default, which entails that the gun became unloaded? It is a prediction problem because when we write the axioms the way Hanks and McDermott wrote them, it is unclear what prediction ought to be made.

But the frame problem, as McDermott correctly points out, was never supposed to be a prediction problem. It was the problem of representing concisely a situation in which the effects of action were known. There is a dual of the Yale shooting problem which is the analogue of the frame problem. Given that we want to prefer the situation in which Fred dies after shooting, after waiting, after loading, how can this preference be represented concisely? There is moreover a non-monotonic character to this preference. If we assert that the gun was in fact unloaded, we want to conclude that Fred is alive, through persistence; we do not want inconsistency, and we do not want to quibble over Fred's living. If we then assert that Fred is dead, this should be possible again without inconsistency.

Two solutions to Yale shooting are recognized. One solution introduces a preference over persistence rules, or over extensions. The other solution redescribes the situation in another ontology, in which there is no issue of preference.

This is the problem that transfixed the knowledge representation and reasoning community for two years. Any solution to the frame problem that anticipated Yale shooting must be praised. How do we avoid "all those damn frame axioms?" A right answer would say, too, how to avoid hacking in those damn preferences among Yale shooting's default rules.

So back to the scene of the crime. Who survived? Begin with the easy cases.

McDermott says of the frame problem that the answer is to use non-monotonic logic. But Yale shooting apparently requires a very different functionality in non-monotonic language: namely, it must be possible to represent preference or stratification of some kind. Or there must be deft use of existing non-monotonic language, in

which ontology has been chosen carefully. In any case, Yale shooting took McDermott broadside and nearly separated him from his non-monotonic soul. McDermott: *dead on arrival.*

Dennett thinks that the frame problem has something to do with "finding the relevant needle of knowledge in the haystack," or with making sure *R2D1* doesn't spend its time proving that various deductions would be irrelevant to its task. Glymour echoes the interpretation. He sees in AI a computational constraint on "old problems" "familiar to philosophers." Representing knowledge? That's supposed to be just like inventing logical languages -- and oh yeah -- there's this morass of mundane sentences we have unexpedtedly to write down too.

The problem of trading control at the object level and control at the meta-level, and doing this all in limited time, is a familiar old problem to AI folk. It is clearly not the frame problem. Even if we could compute with the multitude of frame axioms, the frame problem is unsolved because who would want to write them all down in the first place?

What Dennett in fact identified is a problem that is this year's darling problem in Palo Alto, namely, the problem of limited rationality. This is the problem of how a meta-decision can guide us to the optimal decision, if we need a meta-meta-decision to find the optimal meta-decision. Dennett can rest assured that his essay anticipated *some* important problem in AI, though that problem had nothing to do with the frame problem or with Yale shooting.

II

We can divide the remaining authors into two categories: those who think the frame problem has something to do with ontology, but who are willing to remain "broadly linguisitic" in their representations, and those who believe that the frame problem is the result of choosing a linguistic instead of a pictorial representation. Those with an ontological bent include Janlert, Hayes, and Fodor. Those who would be "quasi-pictorial" include the Dreyfuses and Haugeland.

Hayes survived Yale shooting because he has a firm, simple, unadorned belief that the right ontology will obviate the need for frame axioms. To be sure, Hayes placed his faith in the histories idea, and I have yet to see a treatment of the Yale shooting problem that makes use of histories. But his basic tenet is that we have to "carve nature at the right ontological joints." So hallelujah to the later Lifschitz, to Haugh, Weber, Ginsberg and Baker,

Morgenstern and Stein, and perhaps to my own attempt at introducing explained and unexplained events into the ontology. (A. Baker and M. Ginsberg. "Temporal projection and explanation," *Proc. IJCAI*, 1989. B. Haugh. "Simple causal minimizations for temporal persistence and projection," *Proc. AAAI*, 1987. V. Lifschitz. "Formal theories of action," *Proc. IJCAI*, 1987. R. Loui. "Response to Hanks and McDermott," *Cognitive Science 11*, 1987. L. Morgenstern and L. Stein. "Why things go wrong: a formal theory of causal reasoning," *Proc. AAAI*, 1988. J. Weber. "A versatile approach to action reasoning," U. Rochester Computer Science Technical Report 237, 1988.) Find the right things to minimize -- namely unexplained events, or unmotivated changes, or whatever you want to call them -- and Yale shooting reduces to non-monotonic reasoning. If we insist that a correct analysis of the frame problem's roots should have anticipated Yale shooting, then Hayes has given a correct analysis. Why all the "damn frame axioms"? Because you haven't got the right ontology. Why all the "hacked-in" preferences among defaults? Because you haven't got the right ontology.

We should consider Fodor next. For all the posturing between Hayes and Fodor, the difference in their content is small. Like Hayes, Fodor thinks that the problem has something to do with what we admit into our language. Since Fodor focuses on predicates, technically, his plea is more generally metaphysical than particularly ontological. But to any good nominalist, that is to any good "*x* who nominalizes," the intent is the same. Nonetheless, I think Fodor caught a few Yale bullets.

Fodor says a little more than Hayes. He says that in order to justify the "sleeping dog strategy," we have to make sure we choose the right predicates. That is, in order to use non-monotonic persistence to avoid frame axioms, we have to avoid predicates like *GRUE*, and *FRIDGEON*. The penalty of admitting such poorly chosen predicates is the myriad exceptions: bad predicates means more change; more change means more causal axioms.

Suppose we pass up PARTICLE in favor of *FRIDGEON;* namely, *x* is a *FRIDGEON* at *t* just in case *x* is a particle at *t* and my fridge is on at *t*. Then we have to write down a large number of axioms for the effects of turning on my refrigerator. Note that it would be no penalty if our choices were so consistently poor that we had to reverse our persistence default to a default of constant Heraclitean change. In one case, we abbreviate by specifying only what changes; in the other, we abbreviate by specifying what doesn't change; in either case, we avoid specifying explicitly the banal.

Talk about which predicates should persist is not a new solution to the frame problem, since it employs the old solution: non-monotonic persistence. But our evaluation of Fodor's analysis is worse in light of Yale shooting. Fodor misses the possibility of two persistence rules coming into conflict, which is just what Yale shooting is about. Suppose we choose two predicates so that persistence applies to both of them in order to minimize our causal specification. What happens when they cannot both persist? That is the question that arises with *ALIVE* and *LOADED* in Yale shooting. We want both to persist in general, but we want to indicate our preference for one over the other when a choice is forced.

Fodor's *FRIDGEON* is just like Goodman's *GRUE*. Usually when someone employs *GRUE* or its variants he is pointing out that induction works on some predicates and not on others. That is, things that have been *GREEN* we take to remain *GREEN*. That's not true of *GRUE*. What is so interesting about *GRUE* is that we *do not* take things that have been *GRUE* to remain *GRUE*. It would be odd to suggest that *GRUE* and *GREEN* are examples of predicates, to both of which induction applies, with a *preference* of one over the other. *GRUE* is supposed to be a predicate on which we do not do any induction: not full induction, not half-hearted induction, not induction in deference to induction on *GREEN*. No induction on *GRUE*, period.

Nothing actually prevents us from saying that we want to do induction on both, with a preference for *GREEN* over *GRUE*. This would be like taking *ALIVE* and *LOADED* both to persist, with a preference of one over the other. So if we know *GREEN* to fail at year 2000, namely, my emerald is perhaps phenomenologically pink or orange or ochre or blue, then I should choose that it is phenomenologically blue because I was co-projecting *GRUE*. Nothing prevents saying that if it can't remain *GREEN*, let it remain *GRUE* instead of changing to pink (*i.e.* remaining *GRINK*). If that's what Fodor had said, then it would have corresponded to Yale shooting. But, as I said, that's not what we are trained to think when we think of *GRUE*, or *mutatis mutandis*, when we think of *FRIDGEON*.

Fodor lives to the extent that all of this may still have something to do with induction and the choice of how to apply persistence, and to what, and when. But minimizing causal specification by choosing the right predicates does not help avoid Yale shooting. Approaches to Yale shooting that alter the ontology actually require more specification than those that "just hack in the damn preference."

III

Janlert makes distinctions between explicit and implicit facts, basic and non-basic facts, and primary and secondary phenomena. The bottom line is that some facts are source, while others are derivative. If we keep as much information derivative as possible, then we have something very much like a minimal description of state, a state vector, and all changes and non-changes of derivative facts come along for the ride for free. This doesn't excuse us from the banal, since the color of a block needs to be made explicit; it can't be derived from absolute position. But it does excuse us from the redundant, such as relative position, which can remain implicit and indeed be derived from absolute position. It doesn't excuse the frame axioms, since painting a block leaves its position unchanged. But it does mimimize the number of frame axioms and causal axioms required. Once we have recorded the effects and non-effects on source facts, we have implicitly recorded the effects and non-effects on derivative facts.

So what of Yale shooting's *ALIVE* and *LOADED*? Should one fluent be source and the other be derivative? Suppose we had a fluent, *SMELLS-BAD*, which is true just in case there's a dead body in the room and a gun has just been fired. *SMELLS-BAD* is certainly derivative. Its truth can be determined biconditionally from the falsity of *ALIVE* and the effect of firing. The truth of *ALIVE* can in a single instance be determined, after firing occurs in the context of *LOADED*. But in general, *ALIVE* cannot be derived from *LOADED*. There is symmetry, too. *LOADED* can in a single instance be derived from *ALIVE*. But it cannot be derived in general from *ALIVE*. Neither seems more basic than the other. I am assuming that Janlert would not want to make *ALIVE*-after-shooting derivative while *ALIVE*-after-waiting is source. This leaves no general method for making the source/derivative distinction.

What Janlert might do is refine his partition. There could be degrees to which something is source or derivative. *LOADED* is more basic than *ALIVE*, which in turn is more basic than *SMELLS-BAD*. In Yale shooting, we attempt to derive the truth or falsity of *ALIVE*, from *LOADED*, and the effect of firing. In case we fail to do so, we take *ALIVE* to be unaffected by the action. We never try to derive the truth or falsity of *LOADED* from *ALIVE*. Thus, we have a stratification solution to Yale shooting. This is a liberal interpretation of Janlert, and would make Janlert a survivor.

More likely, Janlert's original intent is to avoid the frame problem, and hence Yale shooting, by insisting on a purely monotonic representation. What is so appealing about a minimal,

explicit state description? The effects of actions can be specified *in toto*. If only changes and non-changes of explicit facts need be represented, how awful can specifying non-changes be? Thus, the explicit and certain effect of moving a block is that it leaves *color* unchanged. The explicit and certain effect of waiting just is to leave *LOADED* unchanged. This does not have the non-monotonic character we wanted; we cannot add knowledge that the gun was *UNLOADED* after the wait without retraction; this violates an original desideratum. Here, non-monotonicity is superfluous, because properly choosing which facts to make explicit avoids the ugly numerousness of frame axioms. And it becomes impossible even to describe the Yale shooting problem. I am not sure if retreat to monotonicity amounts to Janlert's survival or suicide.

Janlert also mentions pictorial representations. But it is Haugeland who gives the idea real consideration. What is the frame problem at base? To Haugeland, it is the result of insisting on linguistic instead of pictorial representation. Scale models are examples of pictorial representation. In a scale model of the blocks world, we use a small cube to stand for block 17 instead of using the six-character string *BLOCK17*. We can determine the effects of moving blocks quite easily. Decide what in the model is analogous to the particular move action. Perform the analogous move. Then interpret the resulting state of the model.

Of course, it is not the computation with frame axioms that was the bother, as much as it was the specification of frame axioms. We must determine what about models corresponds to frame axiom specification.

I claim it is the validation of the model (*cum* interpretation) that corresponds to the specification of frame axioms. Consider Haugeland's map of *BMW* garages in Saudi Arabia. Suppose there are two *BMW*'s: a *1989 BMW 750iL* and a vintage pre-war *BMW 328*. When the *750* sedan is driven two miles North, it corresponds to moving a dot on the map one inch nearer the top of the page. What happens to the vintage *328*? The frame axiom says that the *328* stays put; its dot should not move. If we had taped the two dots together, the model or its interpretation would be wrong. Either the model did not correspond to reality under the interpretation; or, moving a dot to the top of the page is not correctly interpreted as driving the *750* sedan North. It is interpreted instead as racing the two Bimmers North at the same speed. Model-builders do not need to write down frame axioms. But model-builders must do something that *prima facie* is equally tedious. They must validate the semantically significant side-effects and non-side-effects of the model. They must make sure that the *328*'s dot moves when the *328* moves, and stays put when the *328* stays put.

I say validation is tedious *prima facie*, because no one builds a model and validates it, action by action, effect by effect, to wit, frame axiom by implicit frame axiom. Instead, axioms are adopted and validated *en masse*. There is often a correspondence between the nomological laws governing the model and the nomological character of the domain represented. That's why we build a mechanical model of the blocks world; we know that the laws governing little blocks are the same as the laws governing big blocks. If we had an electrical model instead, its validity would be unclear; we might actually have to check its behavior to see what it represents about the world.

Note that pictorial representations are neither monotonic nor non-monotonic. This is because, as Haugeland points out, the implicit/explicit distinction does not apply to such representations. There is no such thing as a theorem. So there is no such thing as monotonicity in the set of theorems. It is possible to reset a model after simulating an action. This clearly retracts an assertion, but it is unclear whether this is due to an augmentation of putative premises from which the assertion was putatively derived. There are no premises and there is no derivation.

What about the preference of extensions in the Yale shooting problem? The bias for a particular conclusion is represented by choosing and validating a scale model that exhibits that bias. If we prefer the solution that Fred dies after Yale shooting, then we might take our scale model to be the paint shop in the Honda factory in Tennessee. Fill the paint cans with metallic blue paint, wait a few seconds, then spray. The preferred outcome is a metallic blue Honda. If we prefer the solution that Fred is alive after Yale shooting, then we need a different model. We can take as our model a cookie jar in the presence of pilfering pre-schoolers. Fill the cookie jar, wait a few hours, then attempt to reward the kids for not robbing the cookie jar. Funny, somehow, the kids go unrewarded. Because there are no frame axioms, there can be no conflict of frame axioms. Both the frame problem and Yale shooting are reduced to the problem of validating models. Thus, Haugeland escapes Yale shooting unscathed.

IV

Finally we have the Dreyfuses. They are difficult to evaluate in the context of Yale shooting because they are not direct about the frame problem. They say that the frame problem is AI's punishment for "substituting discontinuous, flat descriptions for the human capacity to transform past experiences into a continuous perception

of changing relevance." Apparently, humans use pictorial representations and "direct recognition of simliarity." Humans are not alone; holograms are non-anthropomorphic artifacts exhibiting pictorial representation and direct recognition of similarity.

In as much as the Dreyfuses agree with Haugeland on pictorial representation, they too escape Yale shooting. But there is a complication. The Dreyfuses are not so much interested in building quasi-pictorial representations as they are interested in using the envisionment skill in humans.

They want "direct recognition of simliarity." I think it is fair to consider an adequately trained associative memory, such as a connection network, to mechanize this skill. When input is presented to such a holographic memory engine, what is its output? If a blocks world situation and a move action are described in some appropriate input form, presumably the holographic memory engine produces output that corresponds to a situation in which the colors of blocks persist.

We might wonder how it represents the fact that moving a block usually does not change a block's color. It has to do with training. Past experiences with similar input did not raise the issue of changing color. Or they did so with low frequency. Color change was irrelevant to block movement in similar situations in the past. Add a vat of green paint and the association network conjures up memories of re-colored blocks. The locus of the frame axiom in this mechanism is the set of training instances. Past non-change generalizes to future non-change. It is not clear that the tedium of the frame axioms disappears, but in any case, it becomes dominated by the tedium of presenting training instances.

In the context of Yale shooting, it matters whether the output of the association mechanism is unique. Does it ever equivocate between change and non-change? If it does, then there may be a bias toward one and against the other.

This would easily be reflected in the frequency of favorable training instances for each. If I tell you about the Honda paint shop, you probably envision efficiently painted Hondas. But if I mention that the paint cans might have holes in them, you should envision both results -- painted Hondas and painted floor. Your bias for one or the other depends as much on your experience with spray painting as it depends on the inflection I used in my description of the holes. Therein lies the appropriate representation of Yale shooting. The Dreyfuses survive because even if we can imagine imperfections in the direct perception of similarity, we can imagine how to represent a bias.

V

Surely there is a conclusion to all of this.

We represent the world because we need to forsee the effects of alternative actions. A representation that is easily manipulated and reset, and that predicts correctly, will support hypothetical reasoning. We study the frame problem and Yale shooting because they demonstrate weaknesses of our current representations.

Non-monotonic reasoning solved the frame problem, but did not anticipate Yale shooting. Since Yale shooting consumed so much effort and emotion, we are here to recognize those who gave us proper, though unheeded warning.

In Pylyshyn's strangely-timed collection, authors of quite different persuasion produced their analyses of the frame problem. These authors largely avoided the kinds of structures in which the Yale shooting problem could perniciously be posed. Of this, the authors can be proud. The one exception is McDermott, and intellectual historians may choose to think of the Yale shooting literature as a requiem that is his alone.

To be sure, forcing an interpretation of Yale shooting in each author's analysis raises questions that were left unanswered: could two predicates be projected simultaneously, but with different priorities? Could there be degrees to which facts are derivative? Could a mechanism for direct recognition of similarity equivocate? With each question, we see some blood shed, but no additional fatalities.

The real question is whether the pictorial, holographic, minimally explicit, and ontologically defensible representations do for us what we asked in the first place. It's one thing to survive Yale shooting and quite another to produce easily specified, easily manipulated, flexible and accurate, comprehensible representations that can be implemented on a computer. When we ask this larger question in earnest, that is, whose musings are actually going to help AI, I am afraid we see rather more fatalities.

HOW A FORMAL THEORY OF RATIONALITY CAN BE NORMATIVE

Read at the Second International Workshop on Human and Machine Cognition: Android Epistemology, Dunes, Florida, 1991. Later appeared in *Journal of Philosophy 90*, 1993.

(I am indebted to Catherine McKeen, Ronald Chrisley, Donald Nute, and Amos Tversky for discussion, and to Henry Kyburg and Paul Churchland for encouragement.)

This paper was dedicated to Sean Young's Rachel.

I

Even as Davidson advanced the principle of charity as it relates to attribution of mental attitudes, Davidson acknowledged the difficulty of applying formal theory normatively.(see in particular "Psychology as Philosophy" in *Actions and Events*, Oxford, 1980.) Davidson noted that applying decision theory requires interpretation. One must first describe the agent. Once described, coherence with axioms of the theory mandates how the agent must additionally be described if the agent is rational.

In order for a preference of *A* over *B*, with a preference of *B* over *C*, to mandate a preference of *A* over *C*, the formal symbols "$A > B$" and "$B > C$" must be apropos. That is, the agent's observable behavior, "I prefer Mozart to Bartok" must be interpreted as "$A > B$". When, in addition, "Do you prefer Bartok to Mendelssohn?" is met with a nod and represented "$B > C$", the would-be wielder of normative theory must still determine what behavior would count as accord or discord with the formal sentence "$A > C$". With extreme charity, vigorous ostensible denial in response to "Do you prefer Mozart to Mendelssohn?" might still be written "$A > C$".

Davidson was particularly adept at adapting formal language to the situation. He suggested that since an interview with an agent proceeded serially, the preferences thus elicited could be indexed by times:

$$\text{"}A >_{T1} B\text{"},\ \text{"}B >_{T2} C\text{"},\ \text{"}C >_{T3} A\text{"}.$$

Persistence is not normally required of preferences for rational agents,

if $X >_{T1} Y$ then $X >_{T2} Y$,

so there would be no inconsistency with the axioms. Liberal description of the agent guaranteed the agent's rationality.

Decision theorists who argued for a persistence axiom missed the point. Charity of description could be given in any of a number of ways. Richard Jeffrey's framing objects of value was particularly pernicious. ("Risk and Human Rationality," unpublished manuscript, 1983. Jeffrey later published the paper in *Monist 70*, 1987, in which the section to which I refer was omitted. The paper was also advertised as a paper to given at the Boston Philosophy of Science Series, 1986. Professor Jeffrey's use was slightly different. Also, in subsequent correspondence with him, I have acknolwedged that his use is a bit different (aimed at the sure-thing principle instead of transitivity), and that my use of "pernicious" is the rarer derivation from *pernixis* (nimble), or *per nexis* (destructive), not the use that implies malevolence.) "Mozart" in "I prefer Mozart to Bartok" was a different object of value from "Mozart" in "I prefer Mendelssohn to Mozart". The appropriate symbolization was

"*AvsB > BvsA*", "*BvsC > CvsB*", "*CvsA > AvsC*".

Thus, the language of decision theory can be used to describe the agent's preferences. The description is consistent; it contradicts no combination of preference axioms. The agent is apparently rational.

Remarkably, Davidson's charity arose from exigencies of empirical method: he found it unconscionable to call agents irrational while witholding the benefit of indeterminacy. As far as I can tell, his charity did not result from a crisis of conventionalism, or from pursuit of Quine; it happened in the psychology laboratory. One wonders why Kahneman and Tversky are unable to regard their experimental subjects with Davidsonian kindness. (For example, A. Tversky and D. Kahneman, "Judgement under uncertainty: heuristics and biases," *Science 185*, pp. 1124-1131, 1974.) Generalizing indeterminacy in the use of decision theory to other theories of rationality is straightforward. Perhaps too much recent philosophy has been devoted to problems of attributing beliefs to an agent, where belief must be consistent in some formal logical language, and that logical language purports to embody a theory of rational belief. Not just decision theory, but any formal normative theory is at similar risk. A theory of rationality, expressed as constraints on sets of symbols, has no normative force because it cannot fix the translation of situations into symbols. Constraints on habits of translation must either refer to situations informally, or expose themselves to interpretation at the (meta-)

level at which constraints on translation are formalized.

II

How can a formal theory of rationality be normative? Formal theory provides meaning postulates for a specialized language in which an agent can be discussed. The more easily discussed the agent in this specialized language, the more easily the agent is taken to be rational. I have elsewhere taken this pessimistic view of normative theory. (R. P. Loui, "Theory and computation of uncertain inference and decision," Ph.D. dissertation, University of Rochester, 1988. This view is also taken by P. Thagard and R. Nisbett, "Rationality and charity," *Philosophy of Science 50*, pp. 250-267, 1983.) The view is conventionalist: no theory of rationality has prior privilege. It is not categorical: the agent is not simply rational nor irrational with respect to a theory; instead, the agent is more conveniently discussed in one language, less conveniently in another. This view seeks equilibrium: there is interplay between the choice of language and the application of a chosen language to individuals who are judged by the language.

The more powerful a language for prediction, the less easily the language is wielded. The more parameters and syntactic choices in a language that allow indeterminate interpretation, easy wielding, the less predictive the language.

The tradeoff of charity against predictive power in application of normative theory can be viewed much like the tradeoff between error and predictive power in curve fitting, or more general scientific theory formation (here I differ from Davidson, who thinks of nomology in science and is less willing to believe that natural laws can simply be chosen). Offer little charity, that is, make preferences atemporal and objects of value largely indistinguishable, and the agent's behavior can be modeled in the language by rejecting some observations as error. Offer more charity, that is, allow parameters to proliferate, like a high-degree polynomial through a small data set, and indeterminacy weakens all prediction. A preference is recommended because it is predicted by the best theory about the agent's preferences, according to our conventions regarding best theories, not because coherence with the axioms could not otherwise be attained.

Predictive power is all that remains of normative theory's compulsion. The right to compel derives from the fact that a language matches pre-theoretic intuitions about rationality, given the *de facto* habits of translation of a community of language

users.

III

There is another way in which a formal theory can be compulsory.

Formal theory can be implemented; behavior that arises out of implemented theory is thus guided and compelled by the theory. It is one thing to come upon an agent with a formal theory with no interpretation fixed, with no right to fix an interpretation; it is an entirely other thing to adopt a formal theory, fix an interpretation by right, and thereby implement the theory. In some situations, to apply theory, an interpretation must be adduced. In others, to apply theory, an interpretation can be declared. Situations in which this declaration has force are situations in which the theory has normative force: commitment to an interpretation binds the agent, fetters the agent to norm.

Consider the formal symbol system of chess, and the concept of a forced move in a chess position. Coming upon two gods in an ancient Mediterranean sky, or upon two opposing generals in desert sands, or two gaming undergraduates moving tokens about a lunch table, one might wonder whether an interpretation of their behavior as chess-playing in a forcing position would mandate further behavior: whether the forced move is exhibited.

Taking Athens, or the 101st Airborne Division, or the black lacquer button to be a bishop, a move might be forced which corresponds to behavior that the players do not exhibit. One cannot say that they are not chess-players because the imposed interpretation renders their behavior inconsistent with the rules of chess. One must seek a better interpretation. This is a situation that demands charity. Evidence for an interpretation and inclination toward the interpretation might be superb. But interpreters must be prepared to doubt the evidence: giving a hearing to the hypothesis that they are playing chess, the hypothesis must temporarily be placed at the center of the web of belief.

Consider instead the gods or the generals or the undergraduates who consciously decide to play chess with their Greeks, brigades, and buttons. They seek to abide by the rules of chess; they intend to play chess. It is their right to agree on an interpretation, even if by that interpretation, following the rules of the game would be trivial, or by their interpretation, following the rules would seem taxing. What matters is that it is the implementer's right to declare an interpretation of the symbols. After declaration, the interpretation and the rules of the formal system are binding.

Implementers, in fact, have not just the privilege but also the responsibility of fixing the intended interpretation, at least in their own minds.

Forming the intention to instantiate symbols of a formal system is a contract, wherefrom normative force derives. This contract cannot be declared in a formal language, lest there be regress. This contract is frequently made with oneself, or made among oneselves. Nevertheless, it binds.

As an interpreter, one cannot use a logic to prescribe another's beliefs. The best an interpreter can do is provide analysis, the best scientific analysis, respecting all the conventions of science, to determine theory, interpretation, and prediction at once. As an implementer, instead, one can use a logic to form the belief that *P*, once one decides what is to count as the belief that *P*. An implementer, decided in how the system should be implemented, must alter behavior to abide by that system.

Analyst and subject can communicate, agree on interpretation bilaterally. But failure of the bilateral interpretation sometimes provokes search for a new interpretation, unilaterally if necessary. Some bilateral agreements on interpretation bind. In such cases, the parties to the agreement become implementers.

In both interpretation and implementation, there can be unintended interpretations according to which there is conformance with theory. The chess players may intend to violate the rules as declared, but there could be another way of seeing things, according to which they follow the rules. When theory is wielded interpretively, that accident makes them chess players, like it or not; none has the right to declare a privileged interpretation. Accidental, unintended implementations however are less satisfying. Setting out to implement a formal theory of a brain with a trillion Chinese operating hydraulic pumps, and failing, it is meager consolation to know that one of the trillion Chinese implements the theory of the brain by himself.

The doctrine of privileged access stands in poor stead here. When analyst and subject are one, there could still be an interpretive aspect: self-interpretation is not necessarily fixed interpretation. Upon acknowledging belief in *P*, belief in *Q*, and disbelief in *P & Q*, the self-analyst might balk and reinterpret. We customarily suppose we are in a privileged position to know whether we believe *P*. Are we so sure of privileged access to believing *if P then defeasibly Q*? (I have in mind the formal defeasible reason relation from the logical systems reported in J. Pollock, "Defeasible reasoning," *Cognitive Science 11*, pp. 481-518, 1987, and G. Simari and R. Loui, "A mathematical treatment of

defeasible reasoning and its implementation," *Artificial Intelligence 53*, pp. 125-157, 1992.) The rules of the formal system governing defeasible rules can be altered fancifully; any suredness of self-knowledge of defeasible rules will quickly disappear as the semantics of the rules grows arcane. There are many people in this world who just do not know whether they can be said to be playing three-dimensional chess.

IV

Distinguishing implementation from interpretation is important not only because these days, with artificial intelligence, formal systems are implemented on computer systems. There is, too, a particular kind of formal system for rational belief which forces this distinction. These systems are more concerned with construction than with coherence. They confer warrant as the result of process, and outcome of process is non-deterministic. There may be many different computations that constitute a fair hearing, and they do not all reach the same conclusion. A dialectical approach to knowledge is an example of such a process. (For example, the system of N. Rescher, *Dialectics*, SUNY Buffalo, 1977.) A probabilistic test for determining whether an integer is prime is another. (E. Solovay and V. Strassen, "A fast Monte-Carlo test for primality," *SIAM Journal on Computing 6*, pp. 84-85, 1977.) For these constructive conceptions of rational belief, interpretation is ridiculous.

Dialectic is a process that produces warranted belief by adjudicating a disputation. Arguments and counterarguments are produced until some termination condition obtains. Search for arguments and the order in which arguments are advanced are non-deterministic. One disputation might result in a judgement pro; repeating the disputation, the verdict might be reversed. (I once thought I'd cornered Amos Tversky with the principle of charity as an omen of impossibility for normative theory. Tversky replied that any use of decision theory was "as an argument"; indeterminacy of interpretation would allow other arguments, perhaps counterarguments. At the time I did not appreciate the potential of this response. Clearly Tversky has used normative theory interpretively, through fairly uncharitable arguments. But Tversky, with Glenn Shafer, also stands for constructionism in decision theory: utility assessments of non-lotteries are constructed through analogical arguments, which may have counterarguments, and which presumably are vindicated in a dialectical process: for instance, G. Shafer and A. Tversky,

"Languages and designes for probability judgement," *Cognitive Science 9*, pp. 309-339, 1985, or G. Shafer, "Savage revisited," in D. Bell, et al., *Decision Making*, Cambridge, 1988. This would be a third way of fixing an interpretation: not through scientific method, nor by declaration as implementer, but as the result of a disputation.) Primality testing is a celebrated example of a randomized algorithm. Under certain randomization conditions, a number that cannot be factored after several tries is probably prime: the probability rises exponentially as factorizations are tried. One test for primality might result in rational belief that a number is prime; a subsequent test can reverse the verdict.

In each case, belief is rational because it is constructed via a process, not because the belief coheres with other beliefs. Moreover, the outcome of the process could have been otherwise. Like elections and lotteries, the accident of making is what does the making.

The difference between coherence theories of rationality and process theories, as regards charity, is not merely following rules over time versus cohering with rules at a time. Declaring how dynamic rules are to be interpreted is much the same as declaring how static rules are to be interpreted. Proclaim that button-takes-button is at one time "knight takes knight", then button-taking-button is at a later time "rook takes pawn". This is as easily proclaimed as 'five buttons on the table must include one that is not a "knight"'.

The difference, as regards charity, actually lies in the non-determinism of process. When there is non-determinism, there can be charity even when the interpretation is fixed. Many bad decisions are defended by referring to the process by which they were made: charitable interpretation is blind to whether the process used in the defense of the decision is the same process used in construction of the decision. It is too easily claimed that there was a chance set-up with a bad outcome, when perhaps, there was no chance set-up at all, or not the chance set-up that is claimed.

Claim too easily, after the fact, that a hearing was fair. Losing advocates advanced no arguments because, though they were given the opportunity, there was some probability that their search procedures would find none of the effective arguments that could have been constructed. Apparently, this low-probability outcome was realized. Claim too easily, after the fact, that a primality test was fair: selection of would-be divisors failed to find an actual divisor because there was some probability that this could happen. Interpretation permits the lynch mob, post hoc

rationality, bad random number generators. It is easy to interpret an unusual outcome as the result of bad luck in a process, but warranted nevertheless by that process.

Not so easy, though, is actually conducting the trial, holding the debate, or performing the test. The die actually have to be thrown. The advocates must really be given time. The parties to the process must subject themselves to the non-determinism. The judge can still be biased, the die throw bogus, the defense mute. But, at least if there really is non-determinism, the outcomes of chance events cannot simply be faked, the chance set-ups presumed to exist by charitable interpretation. What is needed is some constraint on the post hoc interpretation of chance set-ups.

V

Anyone who has implemented a formal system knows its normative force, or more precisely, the normative force of the contract to implement as intended. Anyone who has attempted to impose a formal theory of rationality knows indeterminacy. Charity threatens anti-intellectualism among formalists. Why invent beautiful formalism if complex constraint can dissipate in insipid interpretation? Because: we want to *construct* attitudes, not just to *have* them.

SHOULD THOSE WHO EXERCISE THE AUTHORITY OF RULES ALSO KNOW THE CASES?

Read at The Third International Workshop on Human and Machine Cognition: Expertise in Context, Seaside, Florida, May 1993.

I

There are those who exercise the authority of rules: individuals empowered by social organization to impose rule-following on others. There are bureaucrats, officers, administrators, governors, umpires, police, librarians, middle managers, hall monitors, and guards at the desk of the gymnasium.

They can provide, withhold, control, determine, impose fines, write tickets, tell you what you can wear, how you can drive, where you can play, how much noise you can make, what you can say, and (I am afraid to say) sometimes, what you can think.

Society has decided to constrain volition. I am concerned with the people whose job it is to carry out the task: those who are given the rules, then given the authority.

What should they know? They are armed with rules. Should they know something else? Should they know the rationales of the rules? Should they understand the context in which the rule was adopted, the context in which it continues to have mandate? Should they know how to reason with the rules, have ideal deductive abilities? Are rules always enough, or should they know too the cases?

These are questions that could be about the constitution of society and the relation of society to individual, about law, about government. If that is how we regard these questions, they could consume a lifetime's pursuit. But they could also be questions about organization and rationality. They could be questions about language, communication, and cognition.

Lately, there has been a lot of experience, designing languages in which rules can be specified, and trying to write rules in these languages so that they might be followed. Specifically new is the experience of formal languages for specifying the control of computer programs. Automata follow rules. Programmers write rules. Computer scientists study this phenomenon.

Recently, the rules have taken the form that social rules, permissions, and obligations might take. I have in mind not only the application of computers to law, but programming itself. And

I have in mind not only rule-based programming, but all of the work on representations that allow conflict among rules.

Problems arose in these designs that led us to question fundamental conceptions of reasoning and the use of language. There was no volition to be controlled by the rules of computer programmers. Computers are perfectly-rule-following social constituents. Still, certain imperfections are unavoidable, in practice, when communicating rules from the organizer to that which is organized. Even when rules can be followed letter by letter, there is a question of what it means to follow rules *to the letter*.

The answers we reach will not surprise anyone who has studied Anglo-American legal tradition. In fact, tradition be Anglo-American or common, our answers will not surprise any legal scholar, nor any political or social scientist. What is new is the nature of the argument. The experience of writing systems of rules in the best conditions provides a new way of looking at what is possible in social settings, and the picture is far from ideal.

I propose a new debunking of the naive notion that rule-givers just give rules, that rules must be followed without exception, and that rules can be imposed without sensitivity to counterarguments based on rationales and precedents. The cognitive and linguistic reflections here are related to the logico-linguistic thinking that repudiates other prevalent and naive notions: the notion that an argument is a proof, or that reasonable people cannot adhere to rules that are in conflict. Argument is not just deduction. Conflicting rules are jointly tenable if there is a strategy for resolving conflict. Rules do not have force whenever they could possibly be applied, that is, whenever their antecedent conditions hold: in most languages for policies, the counterarguments matter. Some of the people who exercise the authority of rules don't seem to understand the rules of the language of rules.

Our experience in this last quarter century as the most active rule-makers to populate the planet, is that there is much wrong with the naive view.

II

Social institutions seek to produce rules through agreements won of negotiation and deliberation. Sometimes there is residual conflict and this conflict permeates the body of rules. That is the most compelling reason why rules contain conflicts, why the language of rules must allow for conflict, and why reasoning about rules must contain strategies for resolving conflict. Moreover,

the rationales of rules, the reasons for their adoption, appear ultimately to be grounded in cases about which there is agreement. Resolving conflicts of rules based on cases is most directly done by considering the cases.

There are other issues. 1. Social decisions are incomplete. Not all possible contingencies are envisioned by legislatures. Guidance has finite range. So rules are supplemented with cases, and the ability to reason from cases, analogically. 2. People and institutions may simply prefer to adopt policies that admit exceptions. So there can be language that admits defeasible rules. 3. Communication among resource-bounded agents is limited. Instead of conveying a massive decision table, a shorthand is adopted. The shorthand is probably non-monotonic (i.e., contains rules that permit exceptions). So unpacking the shorthand, discovering what the rules say, involves non-deductive reasoning.

None of these issues is purely linguistic.

There is yet a purely linguistic argument for my thesis. Supppose there is a complete edict, all possible contingencies envisioned, the people prefer rules that are demonstrative, and the communication is not limited by time or length. The edict must be conveyed, preserved, made public. The tax code, library rules, manager's memo must all be written.

At this point, any of a number of linguistic theses of meaning could be mentioned, with the names of Wittgenstein, Quine, Feyerabend, H.L.A. Hart, or even S. Fish, and dare I say, J. Derrida.

The rules are a formal system. The terms of formal language are not directly applicable in the world. Some guidance needs to be provided. The applicability of formal terms such as *neat appearance*, *excessive noise*, and *departmental business* is broadly informed by our common practices and conventions in natural language.

Guidance on interpretations cannot take the form of additional rules. This would just augment the formal system. The frontier between formal system and practice would be pushed back, but not eliminated.

This is not a new conundrum. Philosophers of science consider the match of theoretical and observational terms. The functionalist theory of meaning considers how the complexity of formal systems constrains the interpretations of their terms. What is the appropriate guidance when interpreting the formal terms in social edicts? Karl Llewellyn, a legal philosopher, is credited with the view that judicial reasoning takes the form of theory-formation and functionalist interpretation. Apparently, since judges can create

law, or at least have some real social authority to invent interpretations of laws, such form of reasoning is appropriate. But there remains the bureaucrat, highway patrolman, and tax accountant, in whom society ostensibly vests no additional authority. Their interpretations rely on analogies. Analogical reasoning is defeasible. Defeasible reasoning depends on process.

The view of reasoning underlying the design of Edwina Rissland's legal reasoning program CABARET ("Artificial intelligence and law," *Yale Law Journal 99*, 1990), is the view I have in mind. Construct an argument for a proposition by invoking a rule. To do so, show that the relevant statute's conditions are met. For instance, to argue for a *home office deduction*, show that the *primary place of business* condition, among others, is met. To show such a condition, there may be other rules that provide guidance. These seek to explicate formal language through more formal language. But at some point, an analogical or case-based argument must be made. A dancer's home studio should be allowed as a *home office* because there is a precedent in which a photographer's studio space was allowed, and there is a relevant similarity.

The Bayesian medical diagnosis software INTELLIPATH also grounds its terms through cases. A *small lesion* actually is defined by the role that it plays in conditioning probabilities. Small to you may not be small to me, but we know for sure that, observed, it halves a particular probability. To ground the formal term, further rules are not provided. Instead, a laser disk displays on-screen images of what the expert diagnostician considered a *small lesion* when inventing the language and probability distributions. Physicians interacting with this expert system are exptected to draw analogies between the depicted lesion and the lesion here and now.

Rule-based systems can be deployed when the interface between the terms of the system and the world is no special problem: that is, when terms refer to measurable quantities, with well-established conventions for measurement, such as *millimeters*, *degrees*, and *dollars*. The amount of instruction required to interface with the program increases as the domain moves to languages more of qualities than of quantities.

The evidence is actually stronger among the failures than among the successes. Every interesting reasoning policy with which we have wanted to imbue our programs from *castle early in chess* to *be relevant in conversation* to *marry young* to *birds fly* to *guns persist in being loaded*, has required non-deductive, non-demonstrative, defeasible reasoning. Other logics underlying our representations, be they inductive, deontic, linear, paraconsistent, multivalued, fuzzy, or not-yet-named, succeed only

to the extent that they are able to simulate in some small way the interplay of rules and exceptions, argument and counterargument, analogy and countervailing analogy.

III

All of this is unsurprising. Formal systems need interpretations. Unlike the situation in scientific theorizing, inventers of normative formal systems intend particular interpretations. To convey this intent, the authors give examples. So why are our bureaucrats and officials instructed only in their books of rules, and why are they typically unmoved by legitimate counterarguments to their analogies? From *honor thy mother and father* to *this FAX machine not for private use*, some people have not been communicating fully.

There is more, and here is the surprise. Analogy forces defeasibility, and it forces the profound kind of defeasibility that presupposes processes of deliberation. Disputational deliberation is anticipated for every case presented.

Distinguish two attitudes toward non-monotonic or defeasible language: one is that defeasibility is an optional property of syntax; the other is an attitude toward the results of partial computation.

Non-monotonic language can be useful merely because it expresses regularities compactly. Some users of non-monotonic language could forgo the convenience of compact specification and return to purely deductive settings. (One is reminded here of Craig's famous interpolation lemma for theoretical and observational terms in scientific language.)

For example: Closed-world assumptions can be made explicit (Reiter, "On closed-world databases," in *Readings in Artificial Intelligence*, Webber and Nilsson, eds., Morgan Kaufman, 1981). *All I know is the flight from St. Louis to Atlanta.* I mean: *There is no flight from St. Louis to Pensacola, no flight from Urbana to Pensacola, no flight from Arcola to Pensacola.* Revision is a bit more involved, but that is separate issue.

For example: Some people might think that the convenience of the following: *by presumption, you may not park; but red stickers permit parking in red spaces;* can be replaced readily with the less convenient: *red stickers permit parking in red spaces; red stickers do not permit parking in yellow spaces; yellow stickers do not permit parking in red spaces; indeterminate sticker color*

does not permit parking in any spaces.

As a final example, Henry Kyburg allows probability arguments that can be defeated by more specific probability arguments (*Logical Foundations of Statistical Inference*, Reidel, 1974). But this is a formalizer's prerogative of brevity. His defeasibility is an artifice; it does not express an attitude toward non-ideal computation. There is an ideal probability, and rational persons are committed to computing it.

No one has ever imagined that there is an ideal analogy.

Disputation over aptness of analogies can continue as long as there are predicates to express similarities and dissimilarities. Reasoning does not converge on an ideal answer, no matter how much computation is allowed. With a finite number of cases finitely described, perhaps analogies run out, but consider what happens when intuitions on hypothetical cases are allowed, or when meta-argument about the superiority of one argument over another is allowed to be analogical. Analogies proliferate.

An author of a corpus of rules who intends that analogies will be made must intend that meanings will be created through the processes that determine good analogies. Since bad analogies are easily made, the process must be disputational, admitting analogy and counter-analogy. Unlike the examples above, defeasible language is not optional for analogies; the entailments of analogies cannot be rewritten in a purely deductive language; the metaphor of disputation suggested by the defeasible language is real.

No particular set of consequences and commitments, no single decision table, no deterministic entailment, results from an edict conveyed in this form. Even if a single interpretation is intended ex ante, the author knows that he cannot constrain the meanings so particularly. Instead, the author constrains those to whom the rules are spoken, to construct conclusions through a process of disputation that uses rules and cases as inputs. The text is raw material for a debate. What makes a home office *deductible*, a manner of driving *reckless*, a neighbour *covetous*, or a news item *fit to print*, is the fair and efficient process by which the conclusion was constructed, from the facts of the matter, the rules containing terms of art, and the cases that show how rules are applied. There is no right answer until there is a deliberation; the outputs of the right kinds of deliberations are by definition in the right.

Meaning is obviously holistic, because exceptions occur. It is also synthetic and constructive on this view of non-monotonic

language: mathematically, it is a relation, not a function. This second view of defeasibility, I have said in the past, is optional for designers of language ("Ampliative inference, computation, and dialectic," in *Philosophy and AI*, Cummins and Pollock, eds., MIT Press, 1991). For some, the view is unavoidable. Those who must ground their formal terms in cases are committed to analogies, hence, committed to this view of their language use.

IV

I do not suppose that it is a good idea for persons to be excepting themselves from rules. I have been on the sore side of a few rules, to be sure, but organization has been mainly beneficial in this social constituent's life. There are a lot of individuals with whom we have to live, upon whom we can all agree that some restraint be imposed. Also, I have met the surly officer, contemptuous bureaucrat, teenage desk monitor, orthodox administrator, unyielding Sabbatarian. I am not comfortable suggesting that the nature of language is such that policies cannot be conveyed: that they must be interpreted by those who are in a position to interpret. Circuit and appellate courts interpret policies, and judges are honorable. But if every bureaucrat were permitted, in fact required, to be an adjudicator, honest men would be lynched daily.

The alternative is worse. An institutional or social edict is produced. Someone is given the power to exercise the authority of that edict. But only the rules are followed, and they are followed as if they are deductive. "It says here that if *A* then *C*, and here is a case of *A*. This demonstrates *C*." What if the rules or cases contain grounds for the argument "if *A & B* then not *C*," and this is a case of *A & B*? Following one rule is not following that which society provided, so under what authority can *C* be claimed? One might as well be following some other edict from some other society.

Rules are made to be broken; no; but they are not the exclusive inputs to deliberations, either.

Rule-followers unreceptive to certain kinds of protest are at fault. "Sorry; that's what the rules say. 'Can't discuss this; 'don't have the authority." In truth, here is authority overstepped. A party to a disputation, especially an adjudicator, must admit argument, counterargument, and rebuttal, from all recognized sources. To close discussion, to bring early termination of deliberation, is to adjudicate most aggressively.

The rules say that certain kinds of arguments can be formed, and they have force in the appropriate context, a context which includes opportunity for counterargument. No argument without counterargument. No interpretation without adjudication. No meaning without interpretation.

We know some things about social organization purely from our study of logic and language:

To be insensitive to cases, rationales, and defeasance of rules is to be autocratic. It does not follow the language of the law, the edict of the authority from which power derives. It is a doing of something else, a making of one's own rules, or more literally, a making whatever one will of the rules. It is a violation of what one has been asked, and permitted, by society to do.