The Need for an Epistemology

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

Abstract

It is argued that we cannot build a sophisticated autonomous planetary rover just by implementing sophisticated planning algorithms. Planning must be based on information, and the agent must have the cognitive capability of acquiring new information about its environment. That requires the implementation of a sophisticated epistemology. Epistemological considerations indicate that the rover cannot be assumed to have a complete probability distribution at its disposal. Its planning must be based upon "thin" knowledge of probabilities, and that has important implications for what planning algorithms might be employed.

1 . Introduction

In thinking about the design of autonomous planetary rovers, the emphasis has usually been on the development of appropriate planning technology. I want to urge that we should spread our net more broadly and think about other aspects of agent cognition as well. In particular, a sophisticated autonomous rover is going to need a sophisticated epistemology, and that has important implications for the planning technology as well.

Imagine an autonomous rover exploring the surface of Mars. It is circumnavigating a large mesa. Two thirds of the way around, it is caught in a rock fall that disables its solar panels and its radio antenna. The sun is setting and the rover has to get back to the lander before it is too dark to see. The rover has been on Mars for a week, and in that time it has learned that parts of the surface form a loose powdery quicksand. If the rover gets into that it will flounder, requiring considerable power to extricate itself. The rover has only its limited battery power to get it back to the lander, and the route back is through new territory. It cannot return the way it came and make it back before dark, and if it is caught in the cold Martian night the power needed to maintain minimal operating temperature will drain the battery leaving it stranded on the floor of the Martian desert.

The rover must return to the lander along the most direct route. Fortunately, it has learned to detect distant fields of quicksand by their color and texture, enabling it to plan a route around them. Of course, it is not infallible at this. Sometimes the rover will find itself surrounded on three sides by quicksand it did not detect earlier and have to backtrack. This time the rover cannot afford much backtracking. To make matters worse, it is getting dark, making it increasingly difficult to make out colors and textures. The rover is equipped with flood lights that can illuminate its path, but they change appearances, making it harder to judge the surface, and their continued use will drain the battery.

The situation deteriorates furhter when the rover notices a haze developing in the west of the sort that foreshadows one of the howling Martian sandstorms. In the thin atmosphere, 200 mph winds could drive the sand into every joint and rapidly disable machinery. The rover knows that, faced with the sandstorm, the lander will eventually use its limited mobility to take refuge in one of the canyons, but without radio contact it cannot notify the rover that it is doing so. They had a previous arrangement regarding where the lander would go in an emergency, just to take care of the possibility of radio failure. But they had not foreseen the simultaneous failure of the solar panels. The rover can either make its way to the landing site or to the canyon, but if it guesses wrong about where the lander will be, it will not have enough power to make up for its mistake.

Fortunately, on the basis of what it had learned during its week on Mars, the rover was able to make a reasonable prediction of how long it would be before the sandstorm reached the lander, and estimate that it was more probable than not that the lander would be waiting at the landing site when the rover arrived. Indeed, it was, the rover was able to plug into the lander's power supply and replenish its battery as they jointly took shelter from the storm, and later it was able to repair its solar panel. The Fourth Martian Expedition ended happily.

This fictional narrative illustrates the kinds of problems an autonomous rover must be able to solve. The rover is faced with a planning problem. Furthermore, the problem requires planning in the face of considerable uncertainty. The rover's planning must be decision-theoretic, in the sense that it must take account of the probabilities and utilities or disutilities of various possible outcomes. However, the rover requires more than sophisticated planning capabilities to solve this problem. It also requires various kinds of information to plug into the planning algorithm, and much of this information cannot be built into the rover prior to the mission. Because the properties of the planet are known only imprecisely before the mission, the rover must be able to learn about its new environment once it arrives and use that new knowledge in its planning. The rover must also have more mundane epistemological capabilities to enable it to make judgments about its current situation.

2. Epistemological Capabilities

2.1 Perceptual Judgments

Let us give a partial enumeration of the kinds of epistemological capabilities the rover needs. To make routine judgments about its current situation on the basis of sensor input, it needs to engage in perceptual reasoning. It is tempting to suppose that this can be accomplished by simply building into the rover a set of probabilities regarding various aspects of its situation given particular patterns of sensory stimulation. These probabilities could then be used directly in decision-theoretic reasoning. However, the amount of information that can be dealt with in this way is severely limited. If the input consists of degrees of activation for various sensors in various regions of the corresponding sensory field, it is hard to imagine an agent that could make sophisticated judgments about its surroundings on the basis of fewer than 300 independent variables. For instance, in the case of vision this would correspond to a visual field of 100 (a 10 by 10 array) of three-colored pixels. It would be impossible to make sensitive visual discrimination with such a visual system. But even for such a severely limited visual system, if all of these 300 variables are relevant to the probability of occurrence of some state *P* of the agent's environment, this requires the agent to encode a probability distribution of 2^{300} independent probabilities. This is approximately 10^{90}. This is an immense number. It has been estimated that the number of elementary particles in the universe is 10^{78}. 2^{300} is twelve orders of magnitude larger, and this is from just 300 relevant variables. Clearly, no agent can store such a probability distribution.

Humans work differently, apparently for good reason. The basic strategy is to take the raw output of the sensors, massage it using various kinds of sophisticated image processing algorithms, and produce a "cognitive output" (as opposed to raw sensor data). What humans are aware of cognitively is a sophisticated visual image already parsed into objects, containing information about relative motion, relative positions of objects, etc. An agent that is cognitively less sophisticated than a human being could simply read off the features of this image as a veridical account of its surroundings. However, human cognition does not regard the image as the epistemological endproduct. If the agent believes the image to be veridical, it will take its surroundings to be the way they appear, but sometimes the agent will bring other non-perceptual knowledge to bear and judge that the image is not a veridical representation of its surroundings. For example, in trying to decide how firm the ground is before it, it is desirable for our rover to be able to decide that the ground is not as red as it appears (redness being an indication of softness we can suppose) because it is being viewed in the light of the setting sun. In other words, the image provides only a defeasible reason for judging that the environment is as represented in the image, and a sophisticated agent should be able to defeat that defeasible presumption and arrive at conflicting beliefs in some cases.[1] So defeasible reasoning from a previously processed image is employed as a computationally more feasible alternative to explicit probabilistic reasoning. To reason in this way, general principles of defeasible reasoning from perceptual images must be implemented in the rover. One implementation of these principles is presented in [13]. It is based on the OSCAR system of defeasible reasoning.

2.2 Temporal Projection

Inferences based upon current perception can provide the rover with some knowledge of its surroundings. However, that is of little use unless

[1] Defeasible reasoning and its role in human knowledge has been studied in philosophy for many years. See Pollock and Cruz [18], Pollock [12], Pollock [15].

the rover can also draw conclusions about its current surroundings on the basis of earlier (at least fairly recent) perception. For instance, suppose the rover wants to judge whether the sand to the right is redder (and hence probably softer) than the sand to the left. We suppose the rover has the ability to make a visual judgment of the redness of the sand it is currently looking at. But as the rover cannot look at both patches of sand at the same time, it will not be able to make the comparison using only current perception. The rover can look at one patch and draw a conclusion about its redness, but when the rover turns to look at the other patch, it no longer has a percept of the first and so is no longer in a position to hold a justified belief about how red it is *now*. This is a reflection of the fact that perception *samples* bits and pieces of the world at disparate times, and a cognitive agent must be supplied with cognitive faculties enabling it to build a coherent picture of the world out of those bits and pieces. What the rover needs is some basis for believing that the first patch of sand has not changed color in the brief interval since it was inspected. In other words, the robot must have some basis for regarding the color as a *stable property* — one that tends not to change quickly over time. This is provided by a defeasible principle of *temporal projection.* As a first approximation, such a principle might have the form:

> If $t_0 < t_1$, believing P-*at*-t_0 is a defeasible reason for the agent to believe P-*at*-t_1.

In [13], I defended a more sophisticated version of this principle and discussed an implementation of it in OSCAR. It is clear, I think, that some such principle of defeasible inference must be included in the epistemology of our rover.

2.3 Causal Reasoning

Obviously, the rover should be able to reason about the causal consequences of its actions and the causal consequences of exogenous events that it witnesses. This requires, among other things, a solution to the frame problem. Again, an efficient implementation of this reasoning within OSCAR is presented in [13].

2.4 Reasoning about Other Agents

In deciding whether the lander will change position before the rover can reach it, the rover must reason about whether the lander is aware of the approaching sandstorm, and about how the lander will evaluate various possible outcomes, and how probable it will take them to be. In other words, the rover must be able to reason about the cognitive state of another agent. This is the robotic problem of other minds.

2.5 Reasoning about Probabilities

Most of the rover's reasoning about its current situation will be based in part on its general beliefs about its environment. Most of these general beliefs will assign conditional probabilities to various eventualities. A certain amount of this probabilistic information can be supplied by the mission designers, but if the world being explored is sufficiently unknown to warrant exploration, then there must be a lot of antecedently unknown probabilities. The rover must have the ability to discover, for example, that the color and texture of the sand is a probabilistic indicator of its softness. The next section will look more carefully at the epistemological problem of acquiring the desired probabilistic knowledge.

2.6 AI and Philosophy

These epistemological problems are problems of philosophy. Some AI researchers have it in their heads the philosophy is something fuzzy and unscientific that serious researchers should avoid. But building an autonomous rover requires taking these problems seriously and implementing solutions to them. You cannot solve the problems without engaging in philosophical analysis.

3. Probability Distributions

The rover requires knowledge of probabilities both to make judgments about its current surroundings and to engage in decision-theoretic planning about what it should do given those current surroundings. Most work on decision-theoretic planning has pretended that the planning agent has at its disposal a complete probability distribution over all the relevant variables of the problem. In other words, where $P_1,...,P_n$ is any set of propositions relevant to the problem (this includes propositions and their negations), the agent knows the value of **PROB**$(P_1\&...\&P_n)$. This probability distribution should reflect both initial knowledge built into the agent by the system designer and any new knowledge the agent has acquired by subsequent experience of its environment.

It is sometimes supposed that the agent begins

with a complete "a priori" probability distribution provided by its designer, and then an updated distribution reflecting newly acquired knowledge can be constructed by Bayesian updating. I think that the power of Bayesian updating is often exaggerated, but let us leave that aside for the moment. I want to focus on the assumption that the planning agent comes to the planning problem equipped with a complete probability distribution, regardless of where that comes from.

In most real-world contexts, this is a completely unrealistic assumption. For a sophisticated agent operating in a novel environment, pretty much anything *might* be relevant. It is up to the agent to discover what is relevant by discovering what variables affect the probabilities of outcomes. So there cannot be much a priori restriction on what variables are included in the probability distribution. But then it follows that it is impossible for a sophisticated agent operating in an environment of real-world complexity to have such a probability distribution at its disposal. The point is not that the agent could not learn the relevant probabilities. It could not even store them all. The problem is a simple cardinality problem. How many variables must be included in this probability distribution? If we have sufficiently limited aspirations, we may only insist that our agent be able to function in a narrowly circumscribed environment. It is hard to imagine any realistic environment of practical interest with, collectively, fewer than 300 variables that are relevant to the probabilities of some of the outcomes. If we want our agent to be able to deal with novel environments of unrestricted complexity, then the number of variables may be larger by many orders of magnitude. But let's suppose there are just 300 two-valued variables relevant to the planning problem. Then the number of conjunctions that must be assigned probabilities by a complete probability distribution is 2^{300}. As before, this is an immense number, on the order of twelve orders of magnitude larger than the current best estimate of the total number of elementary particles in the universe, and this is from just 300 relevant variables. Clearly, no agent can store a complete probability distribution for such a problem in the form of an explicit assignment of probabilities to each conjunction.

Perhaps there is a more efficient way of storing the probability distribution. For example, could we use a Bayesian net with just 300 nodes? Bayesian nets are only helpful if the nodes are sparsely connected, i.e., if most of the probabilities recorded in the net are statistically independent of most of the nodes. How sparse does the net have to be? Well, even if only one in every trillion (10^{12}) probabilities had to be explicitly recorded in the Bayesian net, that would still leave 10^{78} links, i.e., as many links as there are elementary particles in the universe. There is no way to store such a Bayesian net in a real agent.

Can we realistically suppose that our connections are even sparser — so sparse that it is possible to store the Bayesian net in a real agent? Let's consider an example. This generalizes Kushmerick, Hanks and Weld's [4] "slippery gripper" problem. We are presented with a table on which there are 300 numbered blocks, and a panel of numbered buttons. Pushing a button activates a robot arm which attempts to pick up the corresponding block and remove it from the table. We get 100 dollars for each block that is removed. Pushing a button costs two dollars. The hitch is that half of the blocks are greasy, but we don't know which. Initially each block has an equal probability of being greasy. If a block is not greasy, pushing the button will result in its being removed from the table with probability 1.0, but if it is greasy the probability is only 0.1. We are given 300 chances to either push a button or do nothing. In between, we are given the opportunity to look at the table, which costs one dollar, or do nothing. Looking will reveal what blocks are still on the table, but will not reveal directly whether a block is greasy. What should we do? Humans find this problem terribly easy. Everyone I have tried this upon has immediately produced the optimal plan: push each button once, and don't bother to look at the table. Although humans find this problem easy, I have argued recently [16] that no existing planning algorithm can solve this problem. The probabilities of relevance to this problem cannot be represented by a Bayesian net. The relevant variables are whether a block is on the table (T_i), whether a button has been pushed (P_i), and whether a block is greasy (G_i). It would be natural to include a node for each of these (for each stage of the plan), producing a graph with 6400 nodes, and then linking the nodes as necessary to record the primitive probabilistic connections. However, the resulting graph would not be a Bayesian net, because a Bayesian net must be acyclic. If we include nodes for the greasiness of the blocks, acyclicality fails. The probability of a block being on the table after the corresponding button is pushed is influenced by whether it is greasy, and the probability of its

being greasy given that the button is pushed is influenced by whether it is still on the table. If we do not include nodes for the greasiness of the blocks, then the nodes just concern which blocks are on the table at each stage and which buttons have been pushed. This more restricted set of nodes fails to represent all of our probabilistic information, but it does have the form of a Bayesian net. However, as noted above, the probability of a block being greasy is influenced by what other blocks are on the table, and that in turn affects the probability that the block will still be on the table after its button is pushed. Thus the Bayesian net must encode as primitive every probability of the form **PROB**$(T_i/P_i \,\&\, \Pi_{j\in K} T_j)$ where K is a set of block numbers and $i \notin K$. There are 2^{300} such probabilities, so this Bayesian net cannot be encoded in a real agent. The upshot is that even in rather simple domains of real-world complexity, the use of Bayesian nets may not solve the problem of encoding a complete probability distribution in an agent.

What comes immediately to mind is that this planning problem ought to be solvable by employing some kind of factoring or decomposition algorithm, breaking the problem into separate manageable problems for each block individually. But no generally sound algorithm for performing the decomposition comes readily to mind. Existing decomposition schemes do not solve the problem.

Surely, in planetary exploration, more than 300 variables will be potentially relevant to the probabilities of various outcomes. The upshot is that the rover cannot store a complete probability distribution, or even complete probability distributions for relatively small subsets of variables. It must make do with much spottier knowledge of probabilities, and try to acquire new probability knowledge as it needs it. How is this possible?

4. Thin Knowledge of Probabilities

It seems to be unavoidable that agents capable of sophisticated cognition about the real world must learn their probabilities as they go, and they will never have anything approaching a complete probability distribution. We can put this by saying that they will have "thin" knowledge of probabilities. This is worse than just gappy knowledge. What they don't know will, of necessity, be orders of magnitude greater than what they do know. It is worth noting that this does not set probabilities apart from anything else. Real agents will have thin knowledge of just about everything. Many planning algorithms make the *closed world assumption*, according to which the planning agent knows everything there is to know about its environment, including exactly what things of each sort exist. If we look at real agents in the real world, it is hard to imagine an assumption that is farther from the truth.

Real agents must be able to function with very thin knowledge of the world, and that includes knowledge of probabilities. The world is just too big for an agent to know a significant proportion of all the facts about it. How can agents function with such thin knowledge? This is a problem of supreme importance if we want to design sophisticated agents capable of functioning in the real world. In many cases it is defeasible reasoning that enables human beings to bridge the gaps, in effect assuming that decisions can be made and conclusions drawn on the basis of the limited knowledge they have, and assuming that if they knew more, that would not upset those decisions or conclusions. For example, when we reason inductively we extrapolate from our observations and infer defeasibly that things we have not observed will have the same general properties as things we have observed. Similarly, perceptual reasoning builds in the defeasible assumption that things are they way they appear to be, and temporal reasoning builds in a defeasible assumption that the properties of things do not change very fast (the "common sense law of inertia"). There are many other places in which defeasible reasoning plays a crucial role in human knowledge (see [13]).

When we turn to probabilities, we find that agents must also be able to function on the basis of thin knowledge. This gives rise to two questions. First, where do agents (e.g., human beings) get the limited probability knowledge they do have? Second, how can they get by with so little probability knowledge? I will argue that defeasible reasoning plays a crucial role in the answer to both of these questions.

5. Two Kinds of Probabilities

In addressing questions about how to use probabilities in cognition, it is important to realize that there are major philosophical disputes about the foundations of probability theory. These disputes bear upon the properties of probabilities and how they can be used for cognitive tasks like planning. Furthermore, there is more than one

kind of probability, and the different kinds of probabilities have somewhat different logical and mathematical properties. AI researchers are often well versed in "standard" mathematical probability theory, but not in the philosophical foundations. This can be a major problem, because standard mathematical probability theory is just a version of measure theory, and it is often not clear to what extent its results are applicable to "real" probabilities. For instance, mathematical probability theory assumes that probabilities are countably additive, but that assumption is at least debatable when applied to various kinds of "real" probability. Countable additivity has been rejected by most of the important writers in the foundations of probability theory, including de Finetti [2], Reichenbach [20], Jeffrey [3], Skyrms [22], Savage [21], and Kyburg [5]. Philosophical issues in the foundations of probability have direct relevance to how we can build agents that are able to reason probabilistically.

The most important division in the foundations of probability concern the difference between subjective and objective probabilities. Objective probabilities are supposed to represent objective facts about the way the world is. Subjective probabilities are reports of the degree of belief of a cognizer rather than factual statements about the environment [2,3,19,21,22]. An agent's degrees of belief are said to be *coherent* iff they conform to the probability calculus. Subjectivists use what is known as "the Dutch book argument" [19] to argue that rational degrees of belief must be coherent. This argument shows that if an agent's degrees of belief are not coherent, and the agent is willing to accept any bet that appears favorable in terms of its actual degrees of belief, then it is possible to construct a set of bets such that the agent is guaranteed to lose money on the set no matter what happens. There is an extensive literature on the pros and cons of the Dutch book argument, but I will not go into that here.

I believe that subjective probability theory is subject to overwhelming difficulties, and does not provide an adequate foundation for the probabilistic reasoning that an agent must perform to get around in the world. For a detailed discussion of some of these difficulties, see chapter four of [18] and chapter three of [12]. The simplest reason for rejecting the use of subjective probabilities in agent design is that if, as subjectivists often propose, the only constraint on a rational agent's subjective probabilities is that they be coherent, then an agent can attach absolutely any probability to any contingent proposition as long as the probabilities associated with other propositions are adjusted so that the entire set of probabilities is coherent. The probabilities will be completely insensitive to the way the world is. But if we are going to use probabilities as a guide to action, surely we want them to reflect the way the world is. I admit that this may be an overly quick rejection of subjective probability theory, but space precludes a more extensive discussion.

Most theories of objective probability take there to be an intimate connection between probabilities and frequencies.[2] Relative frequencies relate properties. **freq**[*A*/*B*] is the proportion of *B*'s that are *A*'s. For example, we can talk about the frequency with which patches of sand of a certain color are soft. Where #*A* is the cardinality of the set of all *A*'s, **freq**[*A*/*B*] = #(*A*&*B*)/#*B*. The exact connection between objective probabilities and frequencies is controversial,[3] but at the very least, there is an epistemological connection between them. Observing that the relative frequency of *A*'s in a sample of *B*'s is some number *r* gives us a defeasible reason for thinking that **prob**(*A*/*B*), the probability of an arbitrary *B* being an *A*, is approximately *r*. This inference is based upon a general principle of *statistical induction*, and one of the burdens of a theory of objective probability is to formulate such a principle precisely.

Objective probabilities inferred from relative frequencies have the same logical form as the relative frequencies themselves. That is, they relate properties. **prob**(*A*/*B*) is the probability of an arbitrary *B* being an *A*. This is not the probability of a proposition being true. This is an *indefinite probability*, or a "general" probability. Indefinite probabilities are most naturally formulated using free variables. For example, we might write "the probability of a patch of sand of this color being soft" as **prob**(*x* is soft/*x* is sand of this color).

Inductive reasoning and statistical sampling justify beliefs about indefinite probabilities, but the probabilities needed for decision making are the probabilities that particular propositions are true. For example, our rover might conclude inductively that the probability of sand being soft when it looks a certain way is .7. This is an indef-

---

[2] See chapter one of [10] for a survey of objective probability theories.

[3] My own theory is presented in [10].

inite probability about arbitrary times and places. But in deciding whether to attempt to cross this particular patch of sand, what the rover wants to know is how probable it is that *this very patch of sand* is soft. This is the probability of a particular proposition being true, viz., the proposition that this sand is soft. Such probabilities are *definite* or "single case" probabilities.[4] I will follow the convention of symbolizing indefinite probabilities using **prob** and definite probabilities using **PROB**.

Intuitively, both definite and indefinite probabilities make sense. An objective probability theory must accommodate both. Introspecting our own cognition, it seems pretty clear that scientific or inductive reasoning produces knowledge of indefinite probabilities, and then definite probabilities are inferred by somehow applying the indefinite probabilities to particular cases. This kind of inference is called *direct inference*. I will discuss direct inference a bit more fully below.

It is noteworthy that standard mathematical probability theory is only a theory of definite probabilities, not indefinite probabilities. The basis for mathematical probability theory is Kolmogorov's axioms, and according to those axioms probabilities attach to "events", which are best identified with classes of logically equivalent propositions. Indefinite probabilities, dealing as they do with relations (expressed by open formulas in, e.g., first-order logic) have a richer logical structure than definite probabilities. There are numerous principles that hold for indefinite probabilities but cannot even be expressed in the language of the standard probability calculus. Here are three intuitively plausible ones that were discussed in [10]:

(IND) $\mathbf{prob}(Axy/Bxy \;\&\; y = c) = \mathbf{prob}(Axc/Bxc)$.

(PFREQ) $\mathbf{prob}\big(Ax/Bx \;\&\; \mathbf{freq}[Ay/By] = r\big) = r$.

(PPROB) $\mathbf{prob}\big(Ax/Bx \;\&\; \mathbf{prob}(Ay/By) = r\big) = r$.

None of these principles is even well-formed in the standard probability calculus. This is another reflection of the fact that mathematical probability theory may not have much to do with "real" probabilities.

Note that the free variables occurring in definite probabilities are quite different from the "random variables" occurring in the standard probability calculus. If r is a random variable ranging over patches of sand on Mars, then **PROB**(r is soft) is the probability distribution possibly assigning a different value to the definite probability of each patch of sand being soft. **PROB**(r is soft) does not have a single value. On the other hand, **prob**(x is soft/x is a patch of sand on Mars) has a single value. It is, roughly, the proportion of patches of sand on Mars that we would expect to be soft.

It is remarkable how often definite and indefinite probabilities are confused with one another in AI. For example, imagine a medical diagnosis system based on a Bayesian net. Clearly, the probabilities that go into building the net are general probabilities, i.e., indefinite probabilities. But the conclusions of medical diagnosis are the probabilities that specific patients have particular diseases, i.e., they are definite probabilities. Definite probabilities cannot be derived from indefinite probabilities just on the basis of calculations in the probability calculus. Direct inference is required, and as we will see in the next section, that involves more than mathematical calculation. But all Bayesian nets can do is perform calculations in the probability calculus. So this use of Bayesian nets is mathematically invalid.

A sophisticated autonomous rover is going to have to be able to discover indefinite probabilities describing its environment (e.g., when the surface of the ground looks a certain way it is apt to be soft), employ direct inference to infer definite probabilities about its current situation (e.g., the sand in front of it now is probably soft), and then use the latter in decision-theoretic reasoning about what to do. To implement such reasoning in an agent, we first need precise theories about how statistical induction and direct inference should work. These are epistemological theories governing how to reason about indefinite and definite probabilities.

6. Direct Inference

For decision-theoretic reasoning, an agent must know the probabilities of various outcomes resulting from its performing a specific action here and now. This is a definite probability — not an indefinite probability. For example, if the rover is faced with a patch of reddish sand, it will want to know the probability that if it attempts to drive over this patch of sand, it will become bogged down. What is at issue here is the definite proba-

[4] "Single case" is not entirely appropriate terminology, because the proposition can be as general as we like.

bility — the probability that it will become bogged down if it attempts to drive over this very patch of sand. The rover should only be interested in the indefinite probability of becoming bogged down while driving over an arbitrary reddish patch of sand insofar as that helps it to evaluate the definite probability. The indefinite probability is of only "theoretical" interest. The definite probability is of pressing practical concern.

Although our practical interest is in the definite probabilities, it seems clear that the way we get them is by applying our knowledge of indefinite probabilities to the present circumstances. If the sand has a particular reddish cast, and I know that the probability is .95 of sand of that color being soft, then as long as I don't know anything special about this particular patch of sand that would affect the probability, I will infer that the probability that this patch of sand is soft is .95. In other words, I infer **PROB**(St) = .95 from the facts that (1) **prob**(Sx/Rx) = .95 and (2) Rt. This illustrates that although definite probabilities and indefinite probabilities are different beasts, we get definite probabilities by applying indefinite probabilities to particular situations. A theory of direct inference must explain how this works.

The basic idea behind direct inference was first articulated by Hans Reichenbach [20]: in determining the probability that an individual c has a property F, we find the narrowest reference class X for which we have reliable statistics and then infer that **PROB**(Fc) = **prob**($Fx/x\in X$). For example, insurance rates are calculated in this way. There is almost universal agreement that direct inference is based upon some such principle as this, although there is little agreement about the precise form the theory should take.[5]

Direct inference proceeds by using what we know or are justified in believing about particular objects in particular situations to instantiate indefinite probabilities. If we are justified in believing G_1c, G_2c, and G_3c and we know that **prob**(Fx/G_1x & G_2x & G_3x) = r, this gives us a reason for believing that **PROB**(Fc) = r. However, such reasoning is subject to a "total evidence" requirement. If we are also justified in believing G_4c, and we know that **prob**(Fx/G_1x & G_2x & G_3x & G_4x) = s ≠ r, then we should infer that **PROB**(Fc) = s. Thus the original inference must be defeasible, and it is defeated by acquiring additional justified beliefs that instantiate different indefinite probabilities.

As a first approximation, we can capture the dynamics of this reasoning using two principles:

(DI) "Gc & **prob**(Fx/Gx) = r" is a defeasible reason for "**PROB**(Fc) = r".

(SDI) "Hc & **prob**(Fx/Gx & Hx) ≠ **prob**(Fx/Gx)" is an undercutting defeater for (DI).[6]

In the preceding example, by (DI) we have a defeasible reason for believing that **PROB**(Fc) = r, and also a defeasible reason for believing that **PROB**(Fc) = s. In the absence of any other defeaters, these two inferences (to incompatible conclusions) would defeat each other "collectively". However, by (SDI), we also have an undercutting defeater for the inference to the conclusion that **PROB**(Fc) = r, so that inference is defeated leaving the inference to the conclusion that **PROB**(Fc) = s undefeated. Thus we get the effect of the total evidence requirement.[7]

Consider a concrete example. Suppose once more that the rover is deciding whether to attempt to drive over a patch of sand, and it wants to know the probability that it is soft. The sand has a particular reddish cast, and the rover knows that the probability of sand of that color being soft is .95. This gives it a defeasible reason for thinking that the probability is .95 that this sand is soft. However, the sand also has a certain texture, and the rover knows that the probability of sand being soft when it has that combination of color and texture is only .6. Then the rover also has a defeasible reason for thinking that the probability is .6 of this sand being soft. This conflict is resolved by (SDI), according to which the latter inference takes precedence over the former because it is based on more information.

This is only a very rough sketch of a theory of direct inference. In particular, more defeaters than just (SDI) are required to make the theory work. I proposed a more extensive theory in [10]. The point I want to make here is just that the

[5] There are, as far as I know, just three theories of direct inference that have been worked out in detail: Kyburg [5], Levi [6], and Pollock [10]

[6] This assumes the taxonomy of defeaters embodied in the OSCAR system of defeasible reasoning. See [12] or [15] for more detail.

[7] I first introduced this way of handling total evidence requirements in [9].

design of a sophisticated autonomous agent requires us to work out the details of some such theory of direct inference, and implement it. The theory sketched here makes essential use of defeasible reasoning, and I think this will be equally true of any adequate theory. So the theory must be implemented on top of a general-purpose defeasible reasoner. The only existent defeasible reasoner that can deal in a general way with propositions expressed in a non-decidable language (e.g., first-order logic) is my own system OSCAR [12,15], so my own current research in this area aims at implementing direct inference within OSCAR.

7. Statistical Induction

For an agent to infer definite probabilities by direct inference, it must first have knowledge of indefinite probabilities. These are obtained by some form of statistical induction, which might be roughly formulated as follows:

> If X is a sample of B's, "**freq**$[A/X] = r$" gives us a defeasible reason for believing that **prob**(Ax/Bx) is approximately r.

To construct a sophisticated autonomous agent, we must make this principle more precise, give a precise account of the defeaters for it, and implement both in the agent.[8] Notice, again, that this is a principle of defeasible reasoning.

8. Implications for Planning

The conclusion I want to draw from these observations is that we cannot build a sophisticated autonomous planetary explorer just by implementing sophisticated planning algorithms. Planning must be based upon information, both probabilistic and non-probabilistic, and a sophisticated agent exploring a new world cannot have all of its knowledge prepackaged. It must have the ability to learn new things about its world and use that information in its planning. This requires the incorporation of a sophisticated epistemology into the agent. The epistemology will consist of principles of defeasible reasoning both about individual non-probabilistic facts and about definite and indefinite probabilities. The implementation of such principles must be done within a general-purpose defeasible reasoner like OSCAR.

Given more time, I would argue that the decision-theoretic planning algorithms themselves must also proceed in terms of defeasible reasoning. I cannot argue this decisively in the space provided here, but I can at least illustrate the point. First, I am convinced that state-space planners (MDP's, POMDP's, etc.) cannot solve the kinds of problems the rover will face. This is because they require a complete probability distribution, and I have argued that our rover must do its planning on the basis of thin knowledge of probabilities. This is illustrated by the slippery blocks problem of section three. It can be shown that cardinality problems make versions of it unsolvable by all existing state-space planners [16].

One alternative might be to use a decision-theoretic generalization of POCL (partial-order causal-link) planning (e.g., Mahinur [7,8], or decision-theoretic analogues of Buridan [4] or B-Prodigy [1]). I described a general approach to such planners in [17]. However, such planners can only work by drawing their conclusions defeasibly and being prepared to withdraw them in the face of new information. Presumably nobody believes that classical POCL planning is going to be adequate for space exploration, but let us begin by considering it anyway. Standard algorithms based on threat detection and resolution assume that the search for threats is terminating. But in a sophisticated agent, the search for threats cannot just survey a precompiled list. The agent will have to reason about threats using its knowledge of the world, and if its reasoning is non-terminating (e.g., first-order or worse), the search for threats will also be non-terminating. This has the consequence that such algorithms will never produce a plan. (In fact, I showed in [14] that if the set of threats is not recursive then the set of planning problem/solution pairs is not recursively enumerable.) The only way to make such a planner work is to make it defeasible — let it assume defeasibly that there are no threats other than those so far discovered, and produce a plan on that assumption. Discovery of a new threat then constitutes a defeater for the defeasible plan-reasoning. It seems likely that decision-theoretic generalizations of POCL planning will run afoul of the same problem. They must handle classical threats as a special case of more general kinds of probabilistic threats. It seems fairly clear that this will once again produce planning problem/solution pairs that are not

[8] I made a stab at this in [10], but that account is not claimed to be complete.

recursively enumerable. Very likely, the only way to make such a planner work is to have it draw its conclusions defeasibly and then be forever alert to the discovery of new threats.

9. Conclusions

My general conclusion is that most aspects of the cognition of a sophisticated autonomous agent capable of space exploration will require a sophisticated epistemology capable of defeasible reasoning. This is at least true of those aspects of cognition that produce much of the information that is fed into the planning algorithm. There is at least some reason to suspect that the planning algorithm itself must also proceed defeasibly. Many of the problems that must be addressed in order to incorporate such an epistemology into an autonomous rover are traditional philosophical problems. But this is not a license to ignore them. The construction of an autonomous rover will require implemented solutions to these problems.

10. Acknowledgments

This work was supported by NSF grant no. IIS-0080888.

References

[1] Blythe, Jim, and Manuela Veloso, "Analogical replay for efficient conditional planning", *AAAI97*.

[2] de Finetti, B., *Theory of Probability*, vol. 1. New York: John Wiley and Sons, 1974.

[3] Jeffrey, Richard, *The Logic of Decision, 2nd edition*, University of Chicago Press, 1983.

[4] Kushmerick, N., Hanks, S., and Weld, D., "An algorithm for probabilistic planning". *Artificial Intelligence* **76** (1995), 239-286.

[5] Kyburg, Henry, Jr., *The Logical Foundations of Statistical Inference*. Dordrecht: Reidel, 1974.

[6] Levi, Isaac, *The Enterprise of Knowledge*. Cambridge, Mass.: MIT Press, 1980.

[7] Onder, Niluger, and Martha Pollack, "Contingency selection in plan generation", *ECP97.*

[8] Onder, Niluger, and Martha Pollack, "Conditional, probabilistic planning: a unifying algorithm and effective search control mechanisms", AAAI 99.

[9] Pollock, John , "Epistemology and Probability", *Synthese* **81** (1983), 231-252.

[10] Pollock, John, *Nomic Probability and the Foundations of Induction*, Oxford University Press 1990. See [11] for a summary of the theory.

[11] Pollock, John ,"The theory of nomic probability", *Synthese* **90** (1992), 263-300.

[12] Pollock, John, *Cognitive Carpentry*, MIT Press, 1995.

[13] Pollock, John, "Perceiving and reasoning about a changing world", *Computational Intelligence* **14** (1998), 498-562.

[14] Pollock, John, "The logical foundations of goal-regression planning in autonomous agents", *Artificial Intelligence* **106** (1999), 267-335.

[15] Pollock, John, "Defeasible reasoning with variable degrees of justification", *Artificial Intelligence* **133** (2002), 233-282. A considerably expanded (and corrected) version of this paper is available at http://www.u.arizona.edu/~pollock.

[16] Pollock, John, "An easy 'hard problem' for decision-theoretic planners". Available at http://www.u.arizona.edu/~pollock.

[17] Pollock, John, "The logical foundations of decision-theoretic planning in autonomous agents", available at http://www.u.arizona.edu/~pollock.

[18] Pollock, John, and Joseph Cruz, *Contemporary Theories of Knowledge*, 2nd edition, Lanham, Maryland: Rowman and Littlefield, 1999.

[19] Ramsey, Frank, "Truth and probability", in *The Foundations of Mathematics*, ed. R. B. Braithwaite. Paterson, NJ: Littlefield, Adams, 1926.

[20] Reichenbach, Hans, *A Theory of Probability.* Berkeley: University of California Press, 1949. (Original German edition 1935)

[21] Savage, Leonard, *The Foundations of Statistics*, Dover, New York, 1954.

[22] Skyrms, Brian, *Causal Necessity*, Yale University Press, New Haven, 1980.