Uncertainty, Rationality, and Agency

UNCERTAINTY, RATIONALITY, AND AGENCY

Wiebe van der Hoek

Reprinted from *Synthese* 144:2 and 147:2 (2005)
Special Section *Knowledge, Rationality & Action*

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-1-4020-4630-8

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Cover design by Vincent F. Hendricks

Printed on acid-free paper

Printed in the Netherlands.

Contents

WIEBE VAN DER HOEK

Foreword

This book collects all the papers that appeared in 2005 in *Knowledge, Rationality and Action* (KRA), a journal published as a special section of *Synthese*, which addresses contemporary issues in epistemic logic, belief revision, game and decision theory, rational agency, planning and theories of action. As such, the special section appeals to researchers from Computer Science, Game Theory, Artificial Intelligence, Philosophy, Knowledge Representation, Logic and Agents, addressing issues in artificial systems that have to gather information, reason about it and then make a sensible decision about what to do next.

It will be clear already from the contents pages, that this book indeed reflects the core of KRA: the papers in this volume address degrees of belief or certainty, and rational agency. The latter has several manifestations: often constraints on the agent's belief, behaviour or decision making. Moreover, this book shows that KRA indeed represents a 'loop' in the behaviour of the agent: after having made a decision, the life of the agent does not end, rather, it will do some sensing or collect otherwise the outcome of its decision, to update its beliefs or knowledge accordingly and make up its mind about the next decision task.

In fact, the chapters in this book represent two volumes of KRA: the first appeared as a regular volume, the second contained a selection of papers that were accepted for the Conference on Logic and the Foundations of the Theory of Games and Decisions (LOFT 2004). I will now give a brief overview of the themes in this book and of the chapters in the regular volume, the papers of the LOFT-volume are briefly introduced in Chapter five of this book.

The first two chapters, THE NO PROBABILITIES FOR ACTS-PRINCIPLE and A LOGIC FOR INDUCTIVE PROBABILISTIC REASONING, deal with probabilistic reasoning: one in the context of deliberating about future actions, or planning, and the other in that of making inductive inferences. Chapter three, RATIONALITY AS CONFORMITY and Chapter eleven, A LOGICAL FRAMEWORK FOR CONVENTION, both describe rational agents that reason about the rationality of other agents. In Chapter three the challenge of the agent is to act in conformance with the other, in Chapter eleven the emphasis is on predicting the other agents' decision. Both chapters give an account of the reciprocal reasoning that such a decision problem triggers, using notions like common knowledge, common belief, common sense, common reasoning and common reasons for belief.

Reasons for belief are also the topic of Chapter four, ON THE STRUCTURE OF RATIONAL ACCEPTANCE: COMMENTS ON HAWTHORNE AND BOVENS. The chapter investigates ways to deal with the contradiction that arises from three simple postulates of rational acceptance for an agent's beliefs. Chapter six, A SIMPLE MODAL LOGIC FOR BELIEF REVISION, and Chapter seven, PROLEGOMENA TO DYNAMIC LOGIC FOR BELIEF REVISION, both give a modal logical account of the dynamics of a rational agent's belief. Chapter six introduces a belief operator for initial belief and one for the belief after a revision, and Chapter seven gives an account of update when we have many grades of belief. Degrees of belief are also the topic of Chapter eight, FROM KNOWLEDGE-BASED PROGRAMS TO GRADED BELIEF-BASED PROGRAMS, PART I: ON-LINE REASONING: here, the beliefs can be updated by the agent, but are also used to guide his decision during execution of a program.

Chapter nine, ORDER-INDEPENDENT TRANSFORMATIVE DECISION RULES and ten, A PRAGMATIC SOLUTION FOR THE PARADOX OF FREE CHOICE PERMISSION, take us back to formalisations of rational agents again. In Chapter nine the authors focus on the representation of a decision problem for such an agent: the agent prefers certain representations over others, and uses transformation rules to manipulate them. In chapter ten, the rational agent is a speaker in a conversation, and the author uses some ideas from the area of 'only knowing' to model certain Gricean maxims of conversation in order to formally analyse free choice permission.

Regarding the first four chapters, in the miniature THE NO PROBABILITIES FOR ACTS-PRINCIPLE, Marion Ledwig addresses this NPA principle as put forward by Spohn: "*Any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts*". Ledwig discusses several consequences of the principle which are relevant for decision theory, in particular for Dutch book arguments and game theory: the NPA-principle is at odds with conditionalising on one's actions (as done in diachronic Dutch books) and the assumption that one will choose rationally and therefore predict one's choices (as done in game theory). Finally, she makes clear that the NPA-principle refers not to past actions or actions of other persons, but rather to actions that are performable now and extend into the future.

Manfred Jaeger proposes A Logic for Inductive Probabilistic Reasoning in the second chapter of this book. In such kind of reasoning, one applies inference patterns that use statistical background information in order to assign subjective probabilities to subjective events. The author sets himself three design principles when proposing a logical language that formalises inductive probabilistic reasoning: expressiveness, completeness and epistemic justifiability. Indeed, the language proposed enables the encoding of complex probabilistic information, and, by putting an elegant semantics based on logarithmic real-closed values to work, a completeness result for the expressive language is obtained. Finally, regarding justifiability, it is the author's aim to model with the inductive entailment relation a well-justified pattern of defeasible probabilistic reasoning, i.e., to use statistical information to refine an already partially formed subjective probability assignment. For this, it is argued, cross-entropy minimisation relative to possible statistical distributions is the adequate formal model.

In Rationality as Conformity Hykel Hosni and Jeff Paris face the following problem: choose one of a number of options, in such a way that your choice coincides with that of a like-minded, but otherwise inaccessible (in particular non-communicating), agent. In other words, our agent has to 'predict' what an other agent would choose, if that other agent were confronted with the same problem, i.e., to make that choice that coincides with our agent. If a unique option in the space of choices would obviously stand out, that will be the object of choice, and if they are all the same, the best our agents could do is randomise. But what to do in intermediate cases, i.e., where the alternatives are not all alike, but only show *some* structure? In the authors' approach, the agent first singles out a number of outstanding options (called a reason), and then takes a random choice from those. They discuss and mathematically characterise three different reasons: the regulative reason (satisfying weak criteria to choose some naturally outstanding elements: an agent not following them would perform 'unreasonable steps'); the minimum ambiguity reason (a procedural approach based on the notion of indistinguishability of the options) and the smallest uniquely definable reason (take the smallest set of options that is definable in a suitable first-order language). These reasons are then compared and discussed with respect to Game Theory and Rationality.

Gregory Wheeler discusses principles for acceptance of beliefs by a rational agent, in his chapter ON THE STRUCTURE OF RATIONAL ACCEPTANCE: COMMENTS ON HAWTHORNE AND BOVENS. He starts off with observing that the following three principles for rational acceptance together lead to a contradiction: (*i*) it is rational to accept a proposition that is very likely to be true; (*ii*) it is not rational to accept a proposition that you are aware is inconsistent; (*iii*) if it is rational to accept A and also to accept A', that it is rational to accept their conjunction $A \wedge A'$. This is for instance illustrated by the *Lottery Paradox*, in which you rationally accept that each ticket i will not be the winning ticket, but still you don't accept that no ticket will be the winner's. Wheeler's approach is *structural* in the sense that it is deemed necessary to have some connectives in the object language in order to express compound rationally accepted formulas, and to define a notion of logical consequence for such formulas. He then argues that any proposal that solves paradoxes as the one mentioned above, should be structural, in order to bring the conflict between the principles (*i*) and (*iii*) to the fore.

MARION LEDWIG

THE NO PROBABILITIES FOR ACTS-PRINCIPLE[1]

ABSTRACT. One can interpret the No Probabilities for Acts-Principle, namely that any adequate quantitative decision model must in no way contain subjective probabilities for actions in two ways: it can either refer to actions that are performable now and extend into the future or it can refer to actions that are not performable now, but will be in the future. In this paper, I will show that the former is the better interpretation of the principle.

1. INTRODUCTION

Spohn (1977, 1978) claims that his causal decision theory is valuable in part for its explicit formulation of a principle used earlier by Savage (1954, 1972) and Fishburn (1964). This principle, henceforth called the "No Probabilities for Acts"-Principle (or the NPA-Principle) is the following: "*Any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts*" (Spohn 1977, 114).[2] Spohn (1978) maintains that the NPA-Principle isn't used in the rational decision theories of Jeffrey (1965) and of Luce and Krantz (1971), and that this lack is the root for the theories' wrong answers in Newcomb's problem, namely taking only one box (cf. Nozick 1969). According to Spohn (1977) this principle is important, because it has implications for the concept of action, Newcomb's problem, the theory of causality, and freedom of will. In a recent paper, Spohn (1999, 44–45) modifies this principle. He postulates that in the case of strategic thinking, that is, in the case of sequential decision making, the decision maker can ascribe subjective probabilities to his future, but not to his present actions without giving a justification for his claim.[3]

I agree with Spohn that the NPA-principle has implications for the concept of action. If the NPA-principle holds, the decision maker has full control over his actions, that is, he assigns a subjective probability of one to the actions he has decided for and a subjective probability of zero to those he has decided against.[4] Furthermore, it

Synthese (2005) 144: 171–180
Knowledge, Rationality & Action 1–10
DOI 10.1007/s11229-004-2010-6

has implications for a theory of causality if one maintains a probabilistic theory of causality as Spohn (1983) himself does.[5] Finally, it has implications for freedom of will, since an implicit condition for the application of the NPA-principle is that the decision maker is free.

In my opinion, the NPA-principle has some additional consequences for Dutch books (cf. Levi 1987) and game theory (cf. Levi 1997, chap. 2). In the case of diachronic Dutch books, the decision maker must conditionalize on his actions, which violates the NPA-principle. With regard to game theory, the assumption of common knowledge of rationality entails that each agent believes he or she will choose rationally. This means that each agent will be predicting and therefore also assigning probabilities to his or her own choice counter to Levi's contention that deliberation crowds out prediction. So if the NPA-principle holds, game theory has to be built on other assumptions.

I claim that the NPA-principle has some other important consequences:

(1) in opposition to causal decision theories[6] and Kyburg's (1980) proposal to maximize properly epistemic utility, evidential decision theories[7] violate the NPA-principle, because the decision maker conditions his credences by his actions in calculating the utility of an action. Jeffrey's (1983) ratificationism shows a similar feature, for the decision maker conditions his credences by his final decisions to perform his actions. Nozick's (1993) proposal of combining various decision principles also disagrees with the NPA-principle by using evidential decision principles. Meek and Glymour (1994) claim that if the decision maker views his actions as non-interventions in the system, he conditions his credences by his actions, so the NPA-principle is violated here, too. Hence if the NPA-principle is valid, the decision theories which violate it provide wrong solutions to some decision problems and therefore should be abandoned.

(2) By means of the NPA-principle the decision maker cannot take his actions as evidence of the states of the world. The decision maker's credence function cannot be modified by the evidence of the actions, since the NPA-principle demands that the decision maker shouldn't assign any credences to his actions. Thus Jeffrey's (1965) logic of decision, which takes actions as evidence of states of the world, cannot be right if the NPA-principle is valid. Other rational decision theories also assert that the decision maker cannot take his

actions as evidence of the states of the world. In Jeffrey's ratificationism (1983), for example, the decision maker takes his decisions, but not his actions as evidence of the states of the world. In Eells' (1981, 1982, 1985) proposal of the common cause, the decision maker's beliefs and wants and not his actions are evidence of the states of the world. In Kyburg's (1980, 1988) proposal of maximizing properly epistemic utility the decision maker doesn't take his free actions as evidence of the states of the world.

(3) Another consequence of the NPA-principle is to favor Savage's (1954, 1972) trinitarianism, distinguishing between acts, states, and consequences, over Jeffrey's (1965) monotheism, where acts, states, and consequences are all events or propositions, and therefore should be treated all alike.

Due to the great number of the NPA-principle's implications, Spohn (1977, 1978) makes his principle more precise, suggests arguments for it (e.g., point (4)), and points out immediate consequences of it (e.g., point (5)):

(1) The NPA-principle refers to future actions of the decision maker.
(2) Credences for actions do not manifest themselves in the willingness to bet on these actions.
(3) The NPA-principle requires that actions are things which are under the decision maker's full control relative to the decision model describing him.
(4) A theoretical reason for the NPA-principle is that credences for actions cannot manifest themselves in these actions.
(5) An immediate consequence of the NPA-principle is that unconditional credences for events which probabilistically depend on actions are forbidden.

Respective objections to these claims are the following (with regard to point (5) no objection came to my mind):

(1) The term future actions is ambiguous; it can either refer to actions that are performable now and extend into the future or it can refer to actions that are not performable now, but will be in the future.
(2) Why could not the decision maker's probability judgments concerning what the decision maker will do be correlated with the decision maker's willingness to bet? There might be some decision makers, however, who have an aversion to betting and therefore might not be willing to put their money where their mouth is. But

if one forces them to do so, they surely would bet in accordance with their probability judgments.

(3) We do not have to claim that $P(a_1|a_1) = 1$ is a necessary and sufficient condition for full control in order to claim that options are under the decision maker's full control, for a_1 could be a state and not an action. Moreover, one might want to object that Spohn conflates issues about what a person can control with questions about probabilities for actions (Joyce 2002).

(4) Even if credences for actions play no useful role in decision making, Spohn has not shown that they play a harmful role in decision making and should therefore be ommitted (Rabinowicz 2002).

In the following I will explain and criticize in detail only point (1), namely that the NPA-principle refers to future actions of the decision maker. I will begin by presenting Spohn's (1978, 72–73) two examples to provide an intuitive motivation for the NPA-principle: If a friend asks me whether I will be coming to a party tonight and if I answer "yes", then this is not an assertion or a prediction, but an announcement, an acceptance of an invitation, or even a promise. Moreover, if a visitor asks me whether I really believe that I will make a certain move in chess, then I will reply that the question is not whether I believe this, but whether I really want this. That is, in general it can be questioned that, in utterances about one's own future actions, belief dispositions with regard to these actions are manifested. Hence, if I decide to perform a particular action, I also believe I will perform that action.

2. THE NPA-PRINCIPLE REFERS TO FUTURE ACTIONS OF THE DECISION MAKER

The NPA-principle does not refer to past actions and actions of other persons, but only to actions which are open to the decision maker in his decision model, that is, to future actions of the decision maker. Yet "future actions" is ambiguous. It can either refer to actions that are performable now and extend into the future or it can refer to actions that are not performable now, but will be in the future.[8] As I understand Spohn, the NPA-principle refers to actions that are performable now and extend into the future,[9] for Spohn (1977, 115) concedes that decision makers frequently have and utter beliefs about their future actions like the following:

(1) "I believe it is improbable that I will wear shorts during the next winter."

Moreover, Spohn (1977, 116) points out that "As soon as I have to make up my mind whether to wear my shorts outdoors or not, my utterance is out of place." That is, as soon as I have to deliberate about wearing my shorts outdoors now, I cannot say anymore "I believe it is improbable that I will wear shorts outdoors now." Thus according to Spohn decision makers should not assign subjective probabilities to actions that are performable now, but extend into the future.

Yet Spohn (1977, 115) wants this utterance to be understood in such a way that it does not express a credence for an action, but a credence for a decision situation:

(2) "I believe it is improbable that I will get into a decision situation during the next winter in which it would be best to wear shorts."

Thus Spohn assumes that the embedded sentences "I will wear shorts during the next winter" and "I will get into a decision situation during the next winter in which it would be best to wear shorts" are logically equivalent, which is not true. For while it might be the case that I will not wear shorts during the next winter, it might happen that I get into a decision situation during the next winter in which it would be best to wear shorts. Moreover, identifying an action with a decision situation seems to be problematical, as these are clearly two different things.

However, if we, despite the logical inequivalence, concede this opinion to Spohn for a while, we can observe that something else goes wrong. Observe:

(3) "I believe it is improbable that I will run 100 meters in 7 seconds during the next year."

According to Spohn this utterance should be reformulated, since it does not express a genuine probability for an action:

(4) "I believe it is improbable that I will get into a decision situation during the next year in which it would be best to run 100 meters in 7 seconds."

Yet while (3) might be true, (4) might be false. With regard to (3) I know because of my bodily constitution it would not matter how much I tried, I never would be able to run 100 meters in 7 seconds, so indeed I believe it is improbable that I will run 100 meters in 7 seconds during the next year. At the same time with regard to (4) it could happen that the Olympic Games were to take place next year and luckily I qualified for the Olympic team of my country, so that I was in a decision situation in which it would be best for me to run 100 meters in 7 seconds. Thus my belief that it is improbable that I will get into a decision situation during the next year in which it would be best for me to run 100 meters in 7 seconds would be false. Hence there might be a belief context in which (3) is true, but (4) is false.

One might want to object that this reformulation makes sense under the assumption that the decision maker knows that he is strong-willed and thus knows that he will only do what he thinks is best to do and therefore believes it to be improbable to get into a decision situation during the next year in which it would be best to run 100 meters in 7 seconds. True – yet not all decision makers have that constitution. Hence this objection does not generalize.

What is the relevance of these insights? Not much, as the NPA-principle only refers to actions that are performable now and extend into the future, his reformulation of actions that are not performable now, but will be performable in the future, is of no relevance for the NPA-principle. One can deny that [(1) and (2)] and [(3) and (4)] are synonyms and still accept the NPA-principle. Furthermore, one can even ask why it is so important for Spohn to find alternative interpretations of (1) and (3)? By putting forth alternative interpretations, Spohn seems to defend the view that, even in the case of actions that are not performable now but will be performable in the future, the decision maker should not assign any subjective probabilities to his actions. But we have just seen that this is not so, that is, Spohn allows the decision maker to assign subjective probabilities to his actions that are performable in the future. Yet, strictly put in Spohn's view, even utterances like (1) don't express a genuine probability for an action, only a probability for a decision situation. Thus Levi's (1997, 80) suggestion turns out to be right, namely that these interpretations are meant to express that the decision maker isn't even able to predict his actions that are not performable now, but will be in the future;[10] if utterances like (1) express a probability for a decision situation and not for an action, then the decision maker is not even able to predict his actions that are not performable now, but will be in the future.

3. CONCLUSION

I have clarified that the NPA-principle refers to actions that are performable now and extend into the future.

NOTES

[1] I would like to thank Andreas Blank, Phil Dowe, Alan Hajek, James Joyce, Isaac Levi, Nicholas Rescher, Teddy Seidenfeld, Wolfgang Spohn, Howard Sobel, and especially two anonymous referees from the *BJPS* and two anonymous referees from *KRA* for very helpful comments and discussion. Errors remain my own. I also would like to thank Elias Quinn for correcting and improving my English. A part of this paper was given as a talk in the Fourth In-House Conference in October 2001 during my visit at the Center for Philosophy of Science, University of Pittsburgh, 2001–2002 (cf. also Ledwig 2001).

[2] Trivial conditional subjective probabilities, like $P(a_1|a_1) = 1$ for an action a_1 or $P(a_2|a_1) = 0$ for two disjunctive actions a_1 and a_2, are not considered (Spohn 1977, 1978).

[3] Spohn is not the only one to defend his principle; the weaker thesis that the decision maker should not ascribe subjective probabilities of one or zero to his actions is widely accepted (cf. Ginet 1962; Shackle 1967; Goldman 1970; Jeffrey 1977, 1983; Schick 1979, 1999; Levi 1986). Even the stronger thesis that the decision maker should not ascribe any subjective probabilities to his actions is defended (Levi 1989, 1997; cf. Gilboa 1999). For a discussion of these issues, have a look at Ledwig (forthcoming). As Levi's and Gilboa's arguments for the stronger thesis (with the exception of Levi's betting argument) differ from Spohn's argument, my criticism of the NPA-principle does not hold for these.

[4] One might object that this implication does not hold, if in $P(a_1|a_1) = 1$ a_1 is a state and not an action, because this is simply a consequence of the calculus of probabilities which is true. So this implication only holds, given that one considers only actions as input and does not consider sequential decision problems in which an action may change its status over time from outcome to action to part of the state of the world (cf. Skyrms 1990, 44).

[5] Which causation theory is the adequate one, I want to leave for another paper.

[6] Gibbard and Harper (1978), Skyrms (1980, 1982, 1984), Sobel (1986), Lewis (1981), and Spohn (1978).

[7] Jeffrey (1965, 1988, 1996) and Eells (1981, 1982, 1985).

[8] Spohn does not distinguish between different kinds of future actions.

[9] The extension into the future can be minimal, but needs to be there. Otherwise one could not speak of future actions anymore.

[10] With this we have discovered a further possible implication of the NPA-principle, namely if the NPA-principle holds, the decision maker is not able to predict his own

actions that are performable now and extend into the future. Yet, that deliberation crowds out prediction has already been widely discussed in the literature (Ginet 1962; Jeffrey 1965, 1977, 1983; Shackle 1967; Pears 1968; Goldman 1970, chapter 6; Schick 1979, 1999; Ledwig forthcoming; Levi 1986, Section 4.3, 1989, 1997; cf. Gilboa 1999; Joyce 2002; Rabinowicz 2002). I deal with these authors and their views in Ledwig (forthcoming); moreover, in Ledwig (forthcoming) I defend the thesis that deliberation and prediction are compatible with each other.

REFERENCES

Eells, E.: 1981, 'Causality, Utility and Decision', *Synthese* **48**, 295–327.

Eells, E.: 1982, *Rational Decision and Causality*, Cambridge University Press, Cambridge.

Eells, E.: 1985, 'Causality, Decision and Newcomb's Paradox', in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, The University of British Columbia Press, Vancouver, pp. 183–213.

Fishburn, P. C.: 1964, *Decision and Value Theory*, Wiley, New York.

Gibbard, A. and W. L. Harper: 1978, 'Counterfactuals and Two Kinds of Expected Utility', in C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*, Vol. 1, Reidel, Dordrecht, pp. 125–162.

Gilboa, I.: 1999, 'Can Free Choice be Known', in C. Bicchieri, R. Jeffrey and B. Skyrms (eds.), *The Logic of Strategy*, Oxford University Press, Oxford, pp. 163–174.

Ginet, C.: 1962, 'Can the Will Be Caused?', *The Philosophical Review* **71**, 49–55.

Goldman, A. I.: 1970, *A Theory of Human Action*, Prentice-Hall Inc., Englewood Cliffs, NJ.

Jeffrey, R. C.: 1965, *The Logic of Decision*, McGraw-Hill, New York.

Jeffrey, R. C.: 1977, 'A Note on the Kinematics of Preference', *Erkenntnis* **11**, 135–141.

Jeffrey, R. C.: 1983, *The Logic of Decision*, 2nd edn., The University of Chicago Press, Chicago London.

Jeffrey, R. C.: 1988, 'How to Probabilize a Newcomb Problem', in J. H. Fetzer (ed.), *Probability and Causality*, Reidel, Dordrecht, pp. 241–251.

Jeffrey, R. C.: 1996, 'Decision Kinematics', in K. J. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds.), *The Rational Foundations of Economic Behaviour*, Macmillan, Basingstoke, pp. 3–19.

Joyce, J.: 2002, 'Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions', manuscript.

Kyburg, H. E.: 1980, 'Acts and Conditional Probabilities', *Theory and Decision* **12**, 149–171.

Kyburg, H. E.: 1988, 'Powers', in W. L. Harper and B. Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, Vol. 2, Kluwer, Dordrecht, pp. 71–82.

Ledwig, M.: 2001, 'Some Elaborations on Spohn's Principle', *Proceedings of the 4th In-House Conference*, 26th–28th of October, 2001, Center for Philosophy of Science, University of Pittsburgh, on-line at .

Ledwig, M.: (forthcoming), 'Deliberation, Prediction, and Knowledge', *Conceptus*.

Levi, I.: 1986, *Hard Choices: Decision Making under Unresolved Conflict*, Cambridge University Press, Cambridge/ New York/ Port Chester/ Melbourne/ Sydney.
Levi, I.: 1987, 'The Demons of Decision', *Monist* **70**, 193–211.
Levi, I.: 1989, 'Rationality, Prediction and Autonomous Choice', *Canadian Journal of Philosophy* Supplemental. Vol. 19, pp. 339–363, Reprinted in I. Levi (1997), pp. 19–39.
Levi, I.: 1997, *The Covenant of Reason: Rationality and the Commitments of Thought*, Cambridge University Press, Cambridge.
Lewis, D.: 1981, 'Causal Decision Theory', *Australasian Journal of Philosophy* **59**, 5–30.
Luce, R. D. and D. H. Krantz: 1971, 'Conditional Expected Utility', *Econometrica* **39**, 253–271.
Meek, C. and C. Glymour: 1994, 'Conditioning and Intervening' *British Journal for the Philosophy of Science* **45**, 1001–1021.
Nozick, R.: 1969, 'Newcomb's Problem and Two Principles of Choice', in N. Rescher, D. Davidson and C. G. Hempel (eds.), *Essays in Honor of Carl G. Hempel*, Reidel, Dordrecht, pp. 114–146.
Nozick, R.: 1993, *The Nature of Rationality*, Princeton University Press, Princeton.
Pears, D.: 1968, 'Predicting and Deciding', in P. F. Strawson (ed.), *Studies in the Philosophy of Thought and Action*, Oxford University Press, Oxford, London/New York, pp. 97–133.
Rabinowicz, W.: 2002, 'Does Practical Deliberation Crowd Out Self-Prediction?', *Erkenntnis* **57**, 91–122.
Savage, L. J.: 1954/1972, *The Foundations of Statistics*, Wiley, New York, Dover.
Schick, F.: 1979, 'Self-Knowledge, Uncertainty and Choice', *British Journal for the Philosophy of Science* **30**, 232–252. Reprinted in P. Gärdenfors and N.-E. Sahlin (eds.), *Decision, Probability, and Utility*, Cambridge University Press, Cambridge 1988, pp. 270–286.
Schick, F.: 1999, 'Surprise, Self-Knowledge and Commonality', in B. Hansson, S. Halldén, W. Rabinowicz and N.-E. Sahlin (eds.), *Spinning Ideas – An Electronic Festschrift for Peter Gärdenfors*, on-line at http://www.lucs.lu.se/spinning.
Shackle, G.L.S.: 1967, *Time in Economics*, North-Holland Publishing Company, Amsterdam.
Skyrms, B.: 1980, *Causal Necessity*, Yale University Press, New Haven, London.
Skyrms, B.: 1982, 'Causal Decision Theory', *Journal of Philosophy* **79**, 695–711.
Skyrms, B.: 1984, *Pragmatics and Empiricism*, Yale University Press, New Haven, London.
Skyrms, B.: 1990, 'Ratifiability and the Logic of Decision', in P. A. French, T. E. Uehling Jr. and H. K. Wettstein (eds.), *Midwest Studies in Philosophy, Vol. XV: Philosophy of the Human Sciences*, University of Notre Dame Press, Notre Dame, IND, pp. 44–56.
Sobel, J. H.: 1986, 'Notes on Decision Theory: Old Wine in New Bottles', *Australasian Journal of Philosophy* **64**, 407–437.
Spohn, W.: 1977, 'Where Luce and Krantz Do Really Generalize Savage's Decision Model', *Erkenntnis* **11**, 113–134.
Spohn, W.: 1978, Grundlagen der Entscheidungstheorie, in *Monographien Wissenschaftstheorie und Grundlagenforschung*, No. 8, Scriptor, Kronberg/Ts.

Spohn, W.: 1983, *Eine Theorie der Kausalität*, Habilitationsschrift, Ludwig-Maximilians-Universität München.
Spohn, W.: 1999, *Strategic Rationality*, Forschungsbericht der DFG-Forschergruppe 'Logik in der Philosophie', University of Konstanz.

University of Santa Cruz
Cowell College
Santa Cruz, CA 95064
U.S.A.
E-mail: mledwig@ucsc.edu

MANFRED JAEGER

A LOGIC FOR INDUCTIVE PROBABILISTIC REASONING

ABSTRACT. Inductive probabilistic reasoning is understood as the application of inference patterns that use statistical background information to assign (subjective) probabilities to single events. The simplest such inference pattern is direct inference: from "70% of As are Bs" and "a is an A" infer that a is a B with probability 0.7. Direct inference is generalized by Jeffrey's rule and the principle of cross-entropy minimization. To adequately formalize inductive probabilistic reasoning is an interesting topic for artificial intelligence, as an autonomous system acting in a complex environment may have to base its actions on a probabilistic model of its environment, and the probabilities needed to form this model can often be obtained by combining statistical background information with particular observations made, i.e., by inductive probabilistic reasoning. In this paper a formal framework for inductive probabilistic reasoning is developed: syntactically it consists of an extension of the language of first-order predicate logic that allows to express statements about both statistical and subjective probabilities. Semantics for this representation language are developed that give rise to two distinct entailment relations: a relation $\models$ that models strict, probabilistically valid, inferences, and a relation $\approx$ that models inductive probabilistic inferences. The inductive entailment relation is obtained by implementing cross-entropy minimization in a preferred model semantics. A main objective of our approach is to ensure that for both entailment relations complete proof systems exist. This is achieved by allowing probability distributions in our semantic models that use non-standard probability values. A number of results are presented that show that in several important aspects the resulting logic behaves just like a logic based on real-valued probabilities alone.

1. INTRODUCTION

1.1. *Inductive Probabilistic Reasoning*

Probabilities come in two kinds: as statistical probabilities that describe relative frequencies, and as subjective probabilities that describe degrees of belief. To both kinds of probabilities the same rules of probability calculus apply, and notwithstanding a long and heated philosophical controversy over what constitutes the proper meaning of probability (de Finetti 1937; von Mises 1951; Savage

Synthese (2005) 144: 181–248
Knowledge, Rationality & Action 11–78
DOI 10.1007/s11229-004-6153-2

1954; Jaynes 1978), few conceptual difficulties arise when we deal with them one at a time.

However, in commonsense or inductive reasoning one often wants to use both subjective and statistical probabilities simultaneously in order to infer new probabilities of interest. The simplest example of such a reasoning pattern is that of *direct inference* (Reichenbach 1949, Section 72; Carnap 1950, Section 94), illustrated by the following example: from

(1) 2.7% of drivers whose annual mileage is between 10,000 and 20,000 miles will be involved in an accident within the next year

and

(2) Jones is a driver whose annual mileage is between 10,000 and 20,000 miles

infer

(3) The probability that Jones will be involved in an accident within the next year is 0.027.

The 2.7% in (1) is a statistical probability: the probability that a driver randomly selected from the set of all drivers with an annual mileage between 10,000 and 20,000 will be involved in an accident. The probability in (3), on the other hand, is attached to a proposition that, in fact, is either true or false. It describes a state of knowledge or belief, for which reason we call it a subjective probability.[1]

Clearly, the direct inference pattern is very pervasive: not only does an insurance company make (implicit) use of it in its computation of the rate it is willing to offer a customer, it also underlies some of the most casual commonsense reasoning ("In very few soccer matches did a team that was trailing 0:2 at the end of the first half still win the game. My team is just trailing 0:2 at halftime. Too bad".), as well as the use of probabilistic expert systems. Take a medical diagnosis system implemented by a Bayesian network (Pearl 1988; Jensen 2001), for instance: the distribution encoded in the network (whether specified by an expert or learned from data) is a statistical distribution describing relative frequencies in a large number of past cases. When using the system for the diagnosis of patient Jones, the symptoms that Jones exhibits are entered as evidence, and the (statistical) probabilities of various diseases conditioned on this evidence are identified with the probability of Jones having each of these diseases.

Direct inference works when for some reference class C and predicate P we are given the statistical probability of P in C, and for some singular object e all we know is that e belongs to C. If we have more information than that, direct inference may no longer work: assume in addition to (1) and (2) that

(4) 3.1% of drivers whose annual mileage is between 15,000 and 25,000 miles will be involved in an accident within the next year

and

(5) Jones is a driver whose annual mileage is between 15,000 and 25,000 miles.

Now direct inference can be applied either to (1) and (2), or to (4) and (5), yielding the two conflicting conclusions that the probability of Jones having an accident is 0.027 and 0.031. Of course, from (1), (2), (4), and (5) we would infer neither, and instead ask for the percentage of drivers with an annual mileage between 15,000 and 20,000 that are involved in an accident. This number, however, may be unavailable, in which case direct inference will not allow us to derive any probability bounds for Jones getting into an accident. This changes if, at least, we know that

(6) Between 2.7 and 3.1% of drivers whose annual mileage is between 15,000 and 20,000 miles will be involved in an accident within the next year.

From (1), (2), and (4)–(6) we will at least infer that the probability of Jones having an accident lies between 0.027 and 0.031. This no longer is direct inference proper, but a slight generalization thereof.

In this paper we will be concerned with inductive probabilistic reasoning as a very broad generalization of direct inference. By inductive probabilistic reasoning, for the purpose of this paper, we mean the type of inference where statistical background information is used to refine already existing, partially defined subjective probability assessments (we identify a categorical statement like (2) or (5) with the probability assessment: "with probability 1 is

Jones a driver whose...”). Thus, we here take a fairly narrow view of inductive probabilistic reasoning, and, for instance, do not consider statistical inferences of the following kind: from the facts that the individuals $jones_1, jones_2, \ldots, jones_{100}$ are drivers, and that $jones_1, \ldots, jones_{30}$ drive less and $jones_{31}, \ldots, jones_{100}$ more than 15,000 miles annually, infer that 30%of drivers drive less than 15,000 miles. Generally speaking, we are aiming at making inferences only in the direction from statistical to subjective probabilities, not from single-case observations to statistical probabilities.

Problems of inductive probabilistic reasoning that go beyond the scope of direct inference are obtained when the subjective input-probabilities do not express certainties

(7) With probability 0.6 is Jones a driver whose annual mileage is between 10,000 and 20,000 miles.

What are we going to infer from (7) and the statistical probability (1) about the probability of Jones getting into an accident? There do not seem to be any sound arguments to derive a unique value for this probability; however, $0.6 \times 0.027 = 0.0162$ appears to be a sensible lower bound. Now take the subjective input probabilities

(8) With probability 0.6 is Jones's annual mileage between 10,000 and 20,000 miles, and with probability 0.8 between 15,000 and 25,000 miles.

Clearly, it's getting more and more difficult to find the right formal rules that extend the direct inference principle to such general inputs.

In the guise of inductive probabilistic reasoning as we understand it, these generalized problems seem to have received little attention in the literature. However, the mathematical structure of the task we have set ourselves is essentially the same as that of *probability updating*: in probability updating we are given a *prior* (usually subjective) probability distribution representing a state of knowledge at some time t, together with new information in the form of categorical statements or probability values; desired is a new *posterior* distribution describing our knowledge at time $t+1$, with the new information taken into account. A formal correspondence between the two problems is established by identifying the statistical and subjective probability distributions in inductive probabilistic inference with the prior and posterior probability distribution, respectively, in probability updating.

The close relation between the two problems extends beyond the formal similarity, however: interpreting the statistical probability distribution as a canonical prior (subjective) distribution, we can view inductive probabilistic reasoning as a special case of probability updating. Methods that have been proposed for probability updating, therefore, also are candidates to solve inductive probabilistic inference problems.

For updating a unique prior distribution on categorical information, no viable alternative exists to *conditioning*: the posterior distribution is the prior conditioned on the stated facts. [2] Note that conditioning, seen as a rule for inductive reasoning, rather than probability updating, is just direct inference again.

As our examples already have shown, this basic updating/inductive reasoning problem can be generalized in two ways: first, the new information may come in the form of probabilistic constraints as in (7), not in the form of categorical statements; second, the prior (or statistical) information may be incomplete, and only specify a set of possible distributions as in (6), not a unique distribution. The problem of updating such partially defined beliefs has received considerable attention (e.g., Dempster 1967; Shafer 1976; Walley 1991; Gilboa and Schmeidler 1993; Moral and Wilson 1995; Dubois and Prade 1997; Grove and Halpern 1998). The simplest approach is to apply an updating rule for unique priors to each of the distributions that satisfy the prior constraints, and to infer as partial posterior beliefs only probability assignments that are valid for all updated possible priors. Inferences obtained in this manner can be quite weak, and other principles have been explored where updating is performed only on a subset of possible priors that are in some sense maximally consistent with the new information (Gilboa and Schmeidler 1993; Dubois and Prade 1997). These methods are more appropriate for belief updating than for inductive probabilistic reasoning in our sense, because they amount to a combination of prior and new information on a more or less symmetric basis. As discussed above, this is not appropriate in our setting, where the new single case information is not supposed to have any impact on the statistical background knowledge. Our treatment of incompletely specified priors, therefore, follows the first approach of taking every possible prior (statistical distribution) into account (see Section 4.1 for additional comments on this issue).

The main problem we address in the present paper is how to deal with new (single-case) information in the form of general probability constraints. For this various rules with different scope of application have previously been explored. In the case where the new constraints prescribe the probability values $p_1, \ldots, p_k$ of pairwise disjoint alternatives $A_1, \ldots, A_k$, *Jeffrey's rule* (Jeffrey 1965) is a straightforward generalization of conditioning: it says that the posterior should be the sum of the conditional distributions given the A_i, weighted with the prescribed values p_i. Applying Jeffrey's rule to (1) and (7), for instance, we would obtain $0.6 \times 0.027 + 0.4 \times r$ as the probability for Jones getting into an accident, where r is the (unspecified) statistical probability of getting into an accident among drivers who do less than 10,000 or more than 20,000 miles.

When the constraints on the posterior are of a more general form than permitted by Jeffrey's rule, there no longer exist updating rules with a similarly intuitive appeal. However, a number of results indicate that *cross-entropy minimization* is the most appropriate general method for probability updating, or inductive probabilistic inference (Shore and Johnson 1980; Paris and Vencovská 1992; Jaeger 1995b). Cross-entropy can be interpreted as a measure for the similarity of two probability distributions (originally in an information theoretic sense (Kullback and Leibler 1951)). Cross-entropy minimization, therefore, is a rule according to which the posterior (or the subjective) distribution is chosen so as to make it as similar as possible within the given constraints to the prior (resp. the statistical) distribution.

Inductive probabilistic reasoning as we have explained it so far clearly is a topic with its roots in epistemology and the philosophy of science rather than in computer science. However, it also is a topic of substantial interest in all areas of artificial intelligence where one is concerned with reasoning and decision making under uncertainty.

Our introductory example is a first case in point. The inference patterns described in this example could be part of a probabilistic expert system employed by an insurance company to determine the rate of a liability insurance for a specific customer.

As a second example, consider the case of an autonomous agent that has to decide on its actions based on general rules it has been programmed with, and observations it makes. To make things graphic, consider an unmanned spacecraft trying to land on some distant planet. The spacecraft has been instructed to choose one of two possible landing sites: site A is a region with a fairly smooth surface, but located in an area subject to occasional severe storms; site B lies in a more rugged but atmospherically quiet area. According to the statistical information the spacecraft has been equipped with, the probabilities of making a safe landing are 0.95 at site A when there is no storm, 0.6 at site A under stormy conditions, and 0.8 at site B. In order to find the best strategy for making a safe landing, the spacecraft first orbits the planet once to take some meteorological measurements over site A. Shortly after passing over A it has to decide whether to stay on course to orbit the planet once more, and then land at A (20 h later, say), or to change its course to initiate landing at B. To estimate the probabilities of making a safe landing following either strategy, thus the probability of stormy conditions at A in 20 h time has to be evaluated. A likely method to obtain such a probability estimate is to feed the measurements made into a program that simulates the weather development over 20 h, to run this simulation, say, one hundred times, each time adding some random perturbation to the initial data and/or the simulation, and to take the fraction q of cases in which the simulation at the end indicated stormy conditions at A as the required probability. Using Jeffrey's rule, then $0.6q + 0.95(1 - q)$ is the estimate for the probability of a safe landing at A.

This example illustrates why conditioning as the sole instrument of probabilistic inference is not enough: there is no way that the spacecraft could have been equipped with adequate statistical data that would allow it to compute the probability of storm at A in 20 h time simply by conditioning the statistical data on its evidence, consisting of several megabytes of meteorological measurements. Thus, even a perfectly rational, automated agent, operating on the basis of a well-defined finite body of input data cannot always infer subjective probabilities by conditioning statistical probabilities, but will sometimes have to engage in more flexible forms of inductive probabilistic reasoning.[3]

1.2. *Aims and Scope*

To make inductive probabilistic reasoning available for AI applications, two things have to be accomplished: first, a formal rule for this kind of probabilistic inference has to be found. Second, a formal representation language has to be developed that allows us to encode the kind of probabilistic statements we want to reason with, and on which inference rules for inductive probabilistic reasoning can be defined.

In this paper we will focus on the second of these problems, basically taking it for granted that cross-entropy minimization is the appropriate formal rule for inductive probabilistic reasoning (see Section 3.1 for a brief justification). The representation language that we will develop is first-order predicate logic with additional constructs for the representation of statistical and subjective probability statements. To encode both deductive and inductive inferences on this language, it will be equipped with two different entailment relations: a relation $\models$ that describes valid probabilistic inferences, and a relation $\approx$ that describes inductive probabilistic inferences obtained by cross-entropy minimization. For example, the representation language will be rich enough to encode all the example statements (1)–(8) in formal sentences $\phi_1, \ldots, \phi_8$.

If, furthermore, ψ_0 is a sentence that says that with probability 0.4 Jones drives less than 10,000 or more than 20,000 miles annually, then we will obtain in our logic

$$\phi_7 \models \psi_0,$$

because ψ_0 follows from ϕ_7 by the laws of probability theory. If, on the other hand, ψ_1 says that with probability at least 0.0162 Jones will be involved in an accident, then ψ_1 does not strictly follow from our premises, i.e.,

$$\phi_1 \wedge \phi_7 \not\models \psi_1.$$

However, for the inductive entailment relation we will obtain

$$\phi_1 \wedge \phi_7 \mid\!\approx \psi_1.$$

Our probabilistic first-order logic with the two entailment relations $\models$ and $\mid\!\approx$ will provide a principled formalization of inductive probabilistic reasoning in an expressive logical framework. The next problem, then, is to define inference methods for this logic. It is well known that for probabilistic logics of the kind we consider here no complete deduction calculi exist when probabilities are required to be real numbers (Abadi and Halpern 1994), but that completeness results can be obtained when probability values from more general algebraic structures are permitted (Bacchus 1990a). We will follow the approach of generalized probabilities and permit probabilities to take values in *logarithmic real-closed fields (lrc-fields),* which provide a very good approximation to the real numbers. With the lrc-field based semantics we obtain a completeness result for our logic. It should be emphasized that with this approach we do not abandon real-valued probabilities: real numbers being an example for an lrc-field, they are, of course, not excluded by our generalized semantics. Moreover, a completeness result for lrc-field valued probabilities can also be read as a characterization of the degree of incompleteness of our deductive system for real-valued probabilities: the only inferences for real-valued probabilities that we are not able to make are those that are not valid in all other lrc-fields. By complementing the completeness result for lrc-field valued probabilities with results showing that core properties of real-valued probabilities are actually shared by all lrc-field valued probabilities, we obtain a strong and precise characterization of how powerful our deductive system is for real-valued probabilities.

The main part of this paper (Sections 2 and 3) contains the definition of our logic $\mathscr{L}_{ip}$ consisting of a probabilistic representation language L_p, a strict entailment relation $\models$ (both defined in Section 2), and an inductive entailment relation $\mid\!\approx$ (defined in Section 3). The basic design and many of the properties of the logic $\mathscr{L}_{ip}$ do not rely on our use of probability values from logarithmic real-closed fields, so that Sections 2 and 3 can also be read ignoring the issue of generalized probability values, and thinking of real-valued probabilities throughout. Only the key properties of $\mathscr{L}_{ip}$ expressed in Corollary 2.11 and Theorem 2.12 are not valid for real-valued probabilities.

To analyze in detail the implications of using lrc-fields we derive a number of results on cross-entropy and cross-entropy minimization in logarithmic real-closed fields. The basic technical results here have been collected in Appendix A. These results are used in Section 3 to show that many important inference patterns for inductive probabilistic reasoning are supported in $\mathscr{L}_{ip}$. The results of Appendix A also are of some independent mathematical interest, as they constitute an alternative derivation of basic properties of cross-entropy minimization in (real-valued) finite probability spaces only from elementary algebraic properties of the logarithmic function. Previous derivations of these properties required more powerful analytic methods (Kullback 1959; Shore and Johnson 1980).

This paper is largely based on the author's PhD thesis (Jaeger 1995a). A very preliminary exposition of the logic $\mathscr{L}_{ip}$ was given in Jaeger (1994a). A statistical derivation of cross-entropy minimization as the formal model for inductive probabilistic reasoning was given in Jaeger (1995b).

1.3. *Previous Work*

Clearly, the work here presented is intimately related to a sizable body of previous work on combining logic and probability, and on the principles of (probabilistic) inductive inference.

Boole (1854) must probably be credited for being the first to combine logic and probability. He saw events to which probabilities are attached as formulas in a (propositional) logic, and devised probabilistic inference techniques that were based both on logical manipulations of the formulas and algebraic techniques for solving systems of (linear) equations (see Hailperin (1976) for a modern exposition of Boole's work).

The work of Carnap (1950, 1952) is of great interest in our context in more than one respect: Carnap was among the first to acknowledge the existence of two legitimate concepts of probability, (in Carnap's terminology) expressing degrees of confirmation and relative frequencies, respectively. The main focus in Carnap's work is on probability as degree of confirmation, which he considers to be defined on logical formulas. His main objective is to find a canonical probability distribution $\mathfrak{c}$ on the algebra of (first-order) formulas, which would allow to compute the degree of confirmation $\mathfrak{c}(h/e)$ of some hypothesis h, given evidence e in a mechanical way, i.e., from the syntactic structure of h and e alone. Such a confirmation function $\mathfrak{c}$ would then be seen as a normative rule for inductive reasoning. While eventually abandoning the hope to find such a unique confirmation function (Carnap 1952), Carnap (1950) proves that for a general class of candidate functions $\mathfrak{c}$ a form of the direct inference principle can be derived: if e is a proposition that says that the relative frequency of some property M in a population of n objects is r, and h is the proposition that one particular of these n objects has property M, then $\mathfrak{c}(h/e)=r$.

Carnap's work was very influential, and many subsequent works on probability and logic (Gaifman 1964; Scott and Krauss 1966; Fenstad 1967; Gaifman and Snir 1982) were more or less directly spawned by Carnap (1950). They are, however, more concerned with purely logical and mathematical questions arising out of the study of probabilistic interpretations for logical language, than with the foundations of probabilistic and inductive reasoning.

In none of the works mentioned so far were probabilistic statements integrated into the logical language under consideration. Only on the semantic level were probabilities assigned to (non-probabilistic) formulas. This changes with Kyburg (1974), who, like Carnap, aims to explain the meaning of probability by formalizing it in a logical framework. In doing so, he develops within the framework of first-order logic special syntactic constructs for statistical statements. These statistical statements, in conjunction with a body of categorical knowledge, then are used to define subjective probabilities via direct inference.

Keisler (1985) and Hoover (1978) developed first-order and infinitary logics in which the standard quantifiers $\forall x$ and $\exists x$ are replaced by a probability quantifier $Px \geqslant r$, standing for "for x with probability at least r". The primary motivation behind this work was to apply new advances in infinitary logics to probability theory.

In AI, interest in probabilistic logic started with Nilsson's (1986) paper, which, in many aspects, was a modern reinvention of (Boole 1854) (see Hailperin (1996) for an extensive discussion).

Halpern's (1990) and Bacchus's (1990a,b) seminal works introduced probabilistic extensions of first-order logic for the representation of both statistical and subjective probabilities within the formal language. The larger part of Halpern's and Bacchus's work is concerned with coding strict probabilistic inferences in their logics. A first approach towards using the underlying probabilistic logics also for inductive probabilistic reasoning is contained in Bacchus (1990b), where an axiom schema for direct inference is presented. Much more general patterns of inductive (or default) inferences are modeled by the *random worlds* method by Bacchus, Grove, Halpern, and Koller (Bacchus et al. 1992, 1997; Grove et al. 1992a,b). By an approach very similar to Carnap's definition of the confirmation function $\mathfrak{c}$, in this method a degree of belief $\Pr(\phi|\psi)$ in ϕ given the knowledge ψ is defined. Here ϕ and ψ now are formulas in the statistical probabilistic languages of Halpern and Bacchus. As ψ, thus, cannot encode prior constraints on the subjective probabilities (or degrees of belief), the reasoning patterns supported by this method are quite different from what we have called inductive probabilistic reasoning in Section 1.1, and what forms the subject of the current paper. A more detailed discussion of