A Dynamic Interaction Between Machine Learning and the Philosophy of Science

JON WILLIAMSON

Department of Philosophy, King's College London, Strand, London WC2R 2LS, UK; E-mail: jon.williamson@kcl.ac.uk

Abstract. The relationship between machine learning and the philosophy of science can be classed as a *dynamic interaction*: a mutually beneficial connection between two autonomous fields that changes direction over time. I discuss the nature of this interaction and give a case study highlighting interactions between research on Bayesian networks in machine learning and research on causality and probability in the philosophy of science.

1. Introduction

How is machine learning related to the philosophy of science?

One view is that machine learning and the philosophy of science are essentially the same subject. Both are concerned with explicating the relation between theory and evidence. Both are concerned with how theories are posited. Both are concerned with how theories are tested. Both are concerned with finding out what makes a theory a good theory.

This is the view taken by Kevin Korb, who stresses 'the fundamental identity between the two disciplines':[1]

> the two disciplines are, in large measure, one, at least in principle. They are distinct in their histories, research traditions, investigative methodologies; however, the knowledge which they both ultimately aim at is in large part indistinguishable.[2]

As evidence for this Korb takes the similarity between meta-learning in machine learning and the search for scientific method in the philosophy of science, and the common concern with accounting for inductive simplicity and theoretical terms. Korb predicts then that 'machine learning and the philosophy of scientific method will coalesce.'[3]

There are, however, fundamental differences between the two disciplines which cast doubt upon this view. First there is a difference in subject matter: while the two disciplines may share common concerns, this is a case of overlap rather than identity. The philosophy of science is a broad-ranging subject, which covers the debate between realism and instrumentalism, the

foundations of quantum mechanics, the units of selection, as well as the nature of causal claims in econometrics. Such topics may have points of relevance to machine learning but are not its subject matter. On the other hand, machine learning has wide scope too – inducing plans for guiding a robot around a room, the analysis of database biases, parsing natural language, optimisation algorithms for search, for example – these are not the topics addressed by philosophers of science.

The second difference is a divergence of aims. Philosophers of science want to understand science – why science is practised the way it is, how it ought to be practised, and what scientific theories tell us about reality, for instance – and this leads to a desire to work within the languages of science, which are predominantly extensions of natural language rather than formal languages, and to favour optimality over practicality, i.e. to develop normative accounts of scientific reasoning that lead to reliable conclusions rather than lead quickly to conclusions. On the other hand machine learning researchers want to develop techniques that allow machines to learn, and this leads to a desire to work with machine languages or formalisms that can easily be implemented, and to favour practicality over optimality, i.e. to develop computationally practical algorithms that may sacrifice optimality.

These differences between the two disciplines motivate the view that the two disciplines are distinct but connected. This is the view taken by Paul Thagard, for example, who envisages a very close relationship between the two subjects:

> Philosophy is continuous with science, differing only in that it deals with issues that are more general, speculative, and normative than those found in individual sciences. The branches of philosophy concerned with reasoning are continuous with psychology and artificial intelligence.[4]

Hilan Bensusan puts the relationship thus:

> Machine learning and philosophy of science are different endeavours: their goals are different in nature. They nevertheless intersect. As they intersect, they can collaborate. With care. It is not a case for a merger. It is rather a case for an affair.[5]

This is also the view that I shall espouse here. I shall argue that there is a two-way relationship between the two distinct disciplines: the philosophy of science can guide machine learning just as machine learning can help the philosophy of science. More particularly I shall argue that there is a *dynamic interaction* between the philosophy of science and machine learning. The concept of dynamic interaction will be explained in Section 2. I will give an overview of evidence for this relationship in Section 3, and examine a case study in closer detail in Section 4.

While a debate about the identity/lack of identity of machine learning and the philosophy of science may call to mind certain obscure heresies that beset the early Christian church,[6] resolving this issue is important for the following reason. If the disciplines are identical it makes sense to merge research institutes that are currently separated. If, however, the disciplines are distinct but dynamically interact then such a merger would be detrimental – it would be more fruitful to proceed independently but with an exchange of ideas.

2. Dynamic Interactions

Donald Gillies and Yuxin Zheng put forward the concept of a dynamic interaction between two disciplines in Gillies and Zheng (2001) and characterised it as follows:

(1) A connection is established between two different fields or subjects (*A* and *B* say).
(2) Both sides may then benefit from this connection, or, more specifically, from the resulting interaction.
(3) The relationship between the two sides is not static but dynamic. Suppose, for example, the flow of ideas is at first principally from *A* to *B*. It will then characteristically happen that, after a while, the direction of the flow of ideas is reversed and ideas from *B* start to influence *A*. Typically the interaction between *A* and *B* has two phases separated by a turning point, although, as is to be expected, the turning point is not precisely defined.
(4) Although the two sides interact, it is important that each one should at the same time preserve some degree of autonomy. As our above analysis of the dynamic nature of their relationship in (3) shows, this is a necessary condition for further development.[7]

In the same paper Gillies and Zheng provide two examples. They posit a dynamic interaction between the philosophy of mathematics and the philosophy of science: the Vienna Circle, who were crucially influenced by philosophy of mathematics, particularly the foundationalism and logical approach of Frege and Russell, invigorated the philosophy of science; later an ailing philosophy of mathematics was lent new life by Lakatos extending Popper's fallibilist philosophy of science to mathematics in his 'Proofs and Refutations', and recent applications of Kuhn's theory of scientific revolutions[8] and Lakatos' methodology of scientific research programmes[9] to mathematics. Their second example is a dynamic interaction between the philosophy of mathematics and computer science: the Turing machine,

the λ-calculus, type theory and first-order logic arose from problems in the foundations of mathematics and were in turn a strong influence on many aspects of computer science; automated theorem proving has led to a renewed interest in the notion of mathematical proof.[10]

A dynamic interaction is, by definition, a good thing for both sides. Such a symbiotic relationship revitalises the two disciplines and ensures their continued survival. An understanding and analysis of the concept of dynamic interaction is also important, for two reasons. The first is retrospective: the existence of dynamic interactions helps explain the success and longevity of a field. The second is prospective: the prospect of dynamic interactions can be used to justify inter-disciplinary research, inter-community discussion and cross-fertilisation of ideas.

3. Overview of the Interaction

In this section I shall give an overview of some of the interactions between the two subjects, many of which are attributable to the clear analogy between hypothesis choice in science and model selection in machines.[11] In Section 4 we shall see in more detail how interactions can help both subjects by looking at a case study.

At the methodological level, the philosophy of science has had a pervasive influence on machine learning. The 1920s saw the establishment of the Vienna Circle study group and the beginning of the rise of logical positivism as a key movement in the philosophy of science. Integral to their approach was the use of logic as a framework for scientific reasoning. The movement was extremely important half way through the twentieth century when Alan Turing put forward his vision of learning machines,[12] and the early machine learning systems of the 1960s were quick to put the learning problem into a logical framework. By the end of the 1960s the positivists were in decline and an aversion to the normative logical approach to science had set in, in favour of the more descriptive positions championed by Kuhn and Lakatos. Some scepticism towards the symbolic approach to AI followed,[13] and progress on the foundations of neural networks allowed an alternative line of development in machine learning. However it was not long before a resurgence of formal methods gripped the philosophy of science, in the shape of Bayesianism. Bayesian philosophy became important in AI in the 1980s and 1990s with the development of probabilistic expert systems and probabilistic frameworks for learning. Now Bayesian techniques for the machine learning of formal models are widespread, as are Bayesian models themselves.[14]

Machine learning has influenced the controversy in the philosophy of science between inductivism and falsificationism: Gillies argues that the success of the GOLEM machine learning program at learning a scientific

hypothesis provides evidence for inductivism.[15] Proponents of falsificationism remain sceptical about machine learning,[16] but further successes in the automation of scientific hypothesis generation may dampen their ardour. Machine learning has also proven invaluable to the philosophy of science as a means of testing formal models of scientific reasoning, including inductive reasoning,[17] abductive reasoning,[18] coherence-based reasoning,[19] analogical reasoning,[20] causal reasoning (Section 4), and theory revision.[21]

4. Case Study: Bayesian Networks

Bayesian networks have been the hotbed of a particularly productive interaction between machine learning and the philosophy of science, largely due to the fact that the two disciplines share a common interest in causality and probability, two notions which are connected in a Bayesian network.

Philosophers of science have been concerned with the relationship between causality and probabilistic independence and dependence since the 1950s. Hans Reichenbach, for example, advocated a version of the *principle of the common cause*, which says that any two probabilistically dependent events, neither of which causes the other, must be effects of one or more common causes, and the two events must be probabilistically independent conditional on these common causes.[22] Patrick Suppes characterised the relationship in terms of probabilistic dependence rather than independence: cause and effect are probabilistically dependent conditional on the effect's other causes.[23] Wesley Salmon subsequently penned an influential critique of both these views.[24] By the 1980s there had been much exploration of the issues, but little consensus as to the exact relationship, and there was little indication that further philosophical consideration of the issue was leading to a resolution.

Then in the 1980s Judea Pearl took Bayesian philosophy of probability and a principle closely related to the principle of the common cause as a starting point for his theory of Bayesian networks.[25] The *causal Markov condition* says that a cause is probabilistically independent of any of its non-effects conditional on its direct causes. A Bayesian network contains two components: a directed acyclic graph that represents the causal relations on a domain of variables, and probability tables in which the probability distribution of each variable conditional on its direct causes is specified. The causal Markov condition is assumed as a link between the causal graph and probability, and under this assumption a Bayesian network determines a joint probability distribution over the variables.[26] Bayesian networks were originally intended as a formalism for causal reasoning in expert systems, but are now also used as a general framework for representing and learning probability functions. In fact, Bayesian networks were very rapidly adopted by the expert system and machine learning communities,[27] and the

question naturally arose as to whether this acceptance was attributable to the causal Markov condition: the success of Bayesian networks might be taken as evidence in favour of the causal Markov condition as a link between causality and probability.

If the causal Markov condition is a valid assumption, the question also arises as to whether the condition can be used to characterise the causal relation itself: a causal structure may in some sense be the smallest structure to satisfy the condition. Indeed machine learning techniques for learning Bayesian networks have been proposed as a way of automating the task of causal discovery,[28] and the success of such techniques can be used to support a probabilistic analysis of causality based on the causal Markov condition. Such considerations have sparked intense philosophical debate on the nature of causality,[29] and it is clear that machine learning has invigorated issues in the philosophy of science.

In sum, issues in the philosophy of causality and probability influenced the development of Bayesian networks, and the machine learning of Bayesian networks in turn has spurred new philosophical debates about the nature of causality and learning causal relationships. Can the philosophy of science offer machine learning any further insights concerning Bayesian networks? I shall now outline some thoughts in this respect – these are more thoroughly documented in Williamson (2004).

The causal Markov condition implies Reichenbach's principle of the common cause, which has been central to the philosophical discussions of the relationship between causality and probability. Indeed the philosophical literature contains many well-known counter-examples to the principle! If these counter-examples are to be taken seriously then the causal Markov condition cannot hold in general – and in fact it turns out that the failures of the causal Markov condition are quite widespread, and that where causal and probabilistic knowledge is incomplete (as it often is in the application of Bayesian networks) the causal Markov condition is unlikely to hold.[30] Thus a Bayesian network is unlikely to accurately represent the physical probability distribution over its domain.

So what explains the success of Bayesian networks? In my view their success may be attributed to the fact that, even though Bayesian networks are an imperfect representation of physical reality, they are the best models available given the limited knowledge at hand. This can be made precise as follows. Bayesian networks contain two types of knowledge: a causal graph and the probability tables. Now according to the objective Bayesian interpretation of probability,[31] an agent should set her belief distribution over the variables in question to the probability function that maximises entropy given the constraints imposed by her background knowledge. (This is the belief distribution which represents background knowledge but which is otherwise maximally non-committal.) But the probability function that maximises en-

tropy given the knowledge in the causal graph and probability tables of a Bayesian network is precisely the probability function determined by the Bayesian network itself.[32] Hence, even though the causal Markov condition may fail and a Bayesian network may not correctly represent the physical probability distribution on its variables, the network is the model one ought to adopt if one's background knowledge is represented by the components of the network. A Bayesian network is the best model given limited knowledge.

If this is the case, then one should adopt a two-stage methodology when employing Bayesian networks: first form a Bayesian network involving causal and probabilistic knowledge to hand (this is the best model from an objective Bayesian point of view); secondly when there is enough domain data refine this Bayesian network using machine learning techniques to better represent a target probability distribution.[33] This allows one to integrate causal background knowledge with machine learning techniques in a unified formalism, and shows how philosophical considerations to do with the objective Bayesian interpretation of probability can have a bearing on machine learning methodology.

By following the story of Bayesian networks we see how the philosophy of science can interact dynamically with machine learning. The Bayesian interpretation of probability inspired the formulation of Bayesian networks. Machine learning techniques were soon developed to learn Bayesian networks and these techniques allow the possibility of testing philosophical analyses of causality. Meanwhile considerations arising from the philosophy of probability motivate a new approach to learning Bayesian networks and integrating background knowledge.[34]

5. Conclusion

The conditions of Gillies and Zheng for a dynamic interaction appear to have been met in the case of the philosophy of science and machine learning: several connections exist between the two disciplines to the benefit of each side, the relationship varies through time, and both sides remain autonomous.

By adding this interaction to those of Section 2 we obtain a chain of interactions from computer science to machine learning – as in Figure 1. Of course several further connections directly link computer science and machine learning, but they are not altogether autonomous disciplines so there is no direct dynamic interaction between them.
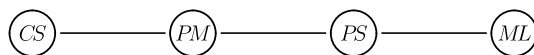

Figure 1. Dynamic interactions: computer science, the philosophy of mathematics, the philosophy of science and machine learning.

This new dynamic interaction may contribute to the success of both the philosophy of science and machine learning, and prospects for future inter-actions can be invoked to encourage proponents of both sides to look for further points of cross-over and pursue inter-disciplinary research.

Acknowledgements

Notes

[1] Korb (2001, p. 1).

[2] Korb (2001, p. 1).

[3] Korb (2001, p. 5).

[4] Thagard (1988, p. 189).

[5] Bensusan (2000, p. 4).

[6] Nestorianism, for example, held that Jesus was two distinct persons, one human and one divine. The Council of Ephesus, convened in 431, concluded that Jesus was one person with two distinct natures, one human and one divine.

[7] Gillies and Zheng (2001, p. 438).

[8] See Gillies (1992).

[9] See Corfield (1997).

[10] See Gillies (2001a) and Barendregt (1997) for further insights into this interaction.

[11] See Bensusan (2000) and Korb (2001) for further discussion of connections between the two disciplines.

[12] Turing (1950, Section 7).

[13] Dreyfus (1972).

[14] The influence of Bayesian philosophy on machine learning has often been indirect, via Bayesian statistics. I am grateful to an anonymous referee for emphasizing this point.

[15] Gillies (1996).

[16] Allen (2001).

[17] Muggleton and de Raedt (1994), Flach (1995).

[18] Dimopoulos and Kakas (1996), Poole (1998), Flach and Kakas (2000), Magnani (2001, Section 4.4).

[19] Thagard (1988, 1992, 2000).

[20] Hofstadter and Group (1995).

[21] Darden (1998).

[22] Reichenbach (1956). Spohn (1980) also advocated an explication of the link between probability and causality in terms of probabilistic independence.

[23] Suppes (1970).

[24] Salmon (1980).

[25] See Pearl (1988, 1993) and also Gillies (2001b) for an account of this development.

[26] See Pearl (1988) or Neapolitan (1990) for the details.

[27] See for example the proceedings of the conferences on Uncertainty in Artificial Intelligence at www.auai.org.

[28] Spirtes et al. (1993), Korb and Nicholson (2003).
[29] See Pearl (1988), Spirtes et al. (1993), McKim and Turner (1997), Glymour and Copper (1999), Hausman (1999), Hausman and Woodward (1999), Korb (1999) and Pearl (2000).
[30] Williamson (2004, Chapter 4).
[31] See Jaynes (2003), Rosenkrantz (1977), William and Corfield (2001).
[32] Williamson (2004, Sections 5.8, 6.1).
[33] See Williamson (2004, Chapter 6) for a more detailed proposal here.
[34] For further applications of Bayesian networks to the philosophy of science see Hartmann and Bovens (2000), Bovens and Hartmann (2000) and Bovens and Olsson (2000).

References

Allen, J.F. (2001), 'Bioinformatics and Discovery: Induction Beckons Again', *BioEssays* 23, pp. 104–107.

Barendregt, H. (1997), 'The Impact of the Lambda Calculus in Logic and Computer Science', *Bulletin of Symbolic Logic* 3(2), pp. 181–215.

Bensusan, H. (2000), 'Is Machine Learning Experimental Philosophy of Science?', in A. Aliseda and D. Pearce, eds., *ECAI2000 Workshop notes on Scientific Reasoning in Artificial Intelligence and the Philosophy of Science*, pp. 9–14.

Boden, M.A. ed. (1990), '*The Philosophy of Artificial Intelligence*', Oxford: Oxford University Press.

Bovens, L. and Hartmann, S. (2000), 'Coherence, Belief Expansion and Bayesian Networks', in *Proc. NMR 2000*, Breckenridge, CO.

Bovens, L. and Olsson, E.J., (2000), 'Coherentism, Reliability and Bayesian Networks', *Mind* 109, pp. 685–719.

Corfield, D. (1997), 'Assaying Lakatos's Philosophy of Mathematics', *Studies in History and Philosophy of Science* 28(1), pp. 99–121.

Corfield, D. and Williamson J. (eds.) (2001), *Foundations of Bayesianism*, *Kluwer Applied Logic Series*, Dordrecht: Kluwer Academic Publishers.

Darden, L. (1998), 'Anomaly-Driven Theory Redesign: Computational Philosophy of Science Experiments', in T.W. Bynum and J.H. Moor, eds., *The Digital Phoenix: How Computers are Changing Philosophy*, New York: Blackwell Publishers, pp. 62–78.

Dimopoulos, Y. and Kakas, A. (1996), 'Abduction and Inductive Learning', in L. De Raedt, ed., *Advances in Inductive Logic Programming*, IOS Press, pp. 144–171.

Dreyfus, H.L. (1972), *What Computers Can't Do*, Cambridge, MA: MIT Press, revised edition 'What Computers Still Can't Do: A Critique of Artificial Reason' 1992.

Flach, P. (1995), 'An Inquiry Concerning the Logic of Induction', *ITK Dissertation Series*.

Flach, P.A. and Kakas, A.C. (2000), 'On the Relation Between Abduction and Inductive Learning', in D.M. Gabbay and R. Kruse, eds., *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 4: Abductive Reasoning and Learning*, Dordrecht: Kluwer Academic Publishers, pp. 1–33.

Gillies, D. (ed.) (1992), *Revolutions in Mathematics*, Oxford: Clarendon Press.

Gillies, D. (1996), *Artificial Intelligence and Scientific Method*, Oxford: Oxford University Press.

Gillies, D. (2001a), 'Logicism and the Development of Computer Science', in A.C. Kakas and F. Sadri, eds., *Computational Logic: Logic Programming and Beyond, Essays in Honour of Robert A. Kowalski*, Lecture Notes in Computer Science 2407, Vol. 2, Berlin: Springer, pp. 588–604.

Gillies, D. (2001b), 'Probability in Artificial Intelligence', in L. Floridi, ed., *The Blackwell Guide to the Philosophy of Computing and Information*, Malden, MA: Blackwell.

Gillies, D. and Zheng, Y. (2001), 'Dynamic Interactions with the Philosophy of Mathematics', *Theoria* 16(3), pp. 437–459.

Glymour, C. and Cooper, G.F. (eds.) (1999), *Computation, Causation, and Discovery*, Cambridge, MA: MIT Press.

Hartmann, S. and Bovens, L. (2000) 'The Import of Auxiliary Theories of the Instrument: A Bayesian-Network Approach', in A. Aliseda and D. Pearce, eds., *ECAI2000 Workshop Notes on Scientific Reasoning in Artificial Intelligence and the Philosophy of Science*, pp. 21–27.

Hausman, D.M. (1999), 'The Mathematical Theory of Causation', *British Journal for the Philosophy of Science* 50, pp. 151–162.

Hausman, D.M. and Woodward, J. (1999), 'Independence, Invariance and the Causal Markov Condition', *British Journal for the Philosophy of Science* 50, pp. 521–583.

Hofstadter, D. and Fluid Analogies Research Group (1995), *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*, Penguin.

Jaynes, E.T. (2003), *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press.

Kakas, A.C. and Sadri, F. (eds.) (2002), *Computational Logic: Logic Programming and Beyond, Essays in Honour of Robert A. Kowalski*, Lecture Notes in Computer Science 2407, Berlin: Springer.

Korb, K.B. (1999), 'Probabilistic Causal Structure', in H. Sankey, ed., *Causation and Laws of Nature*, Dordrecht: Kluwer Academic Publishers, pp. 265–311.

Korb, K.B. (2001), 'Machine Learning as Philosophy of Science', in K. Korb and H. Bensusan, eds., *Proceedings of the ECML-PKDD-01 Workshop on Machine Learning as Experimental Philosophy of Science*, Freiburg.

Korb, K.B. and Nicholson, A.E. (2003), *Bayesian Artificial Intelligence*, London: Chapman and Hall/CRC Press.

Magnani, L. (2001), *Abduction, Reason, and Science: Processes of Discovery and Explanation*, New York: Kluwer Academic/Plenum Publishers.

McKim, V.R. and Turner, S. (1997), *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, Notre Dame: University of Notre Dame Press.

Muggleton, S. and de Raedt, L. (1994), 'Inductive Logic Programming: Theory and Methods', *Journal of Logic Programming* 19, 20, pp. 629–679.

Neapolitan, R.E. (1990), *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*, New York: Wiley.

Pearl, J. (1998), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1993), 'Belief Networks Revisited', *Artificial Intelligence* 59, pp. 49–56.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.

Poole, D. (1998), 'Learning, Bayesian Probability, Graphical Models, and Abduction', in P. Flach and A. Kakas, eds., *Abduction and Induction: Essays on their Relation and Integration*, Dordrecht: Kluwer Academic Publishers.

Reichenbach, H. (1956), *The Direction of Time*, Berkeley and Los Angeles: University of California Press, 1971.

Rosenkrantz, R.D. (1977): *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, Dordrecht: D. Reidel.

Salmon, W.C. (1980), 'Probabilistic Causality', in W.C. Salmon, ed., *Causality and Explanation*, Oxford: Oxford University Press, pp. 208–232.

Salmon, W.C. (1998), *Causality and Explanation*, Oxford: Oxford University Press.

Spirtes, P. Glymour, C. and Scheines, R. (1993), *Causation, Prediction, and Search*, 2nd edition. Cambridge, MA: MIT Press, 2000.

Spohn, W. (1980), 'Stochastic Independence, Causal Independence, and Shieldability', *Journal of Philosophical Logic* 9, pp. 73–99.

Suppes, P. (1970), *A Probabilistic Theory of Causality*, Amsterdam: North-Holland.

Thagard, P. (1988), *Computational Philosophy of Science*, Cambridge, MA: MIT Press/ Bradford Books.

Thagard, P. (1992), *Conceptual Revolutions*, Princeton: Princeton University Press.

Thagard, P. (2000) 'Probabilistic Networks and Explanatory Coherence', *Cognitive Science Quarterly* 1, pp. 91–114.

Turing, A.M. (1950), 'Computing Machinery and Intelligence', in M.A. Boden, ed., *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press, pp. 40–66.

Williamson, J. (2004), *Bayesian Nets and Causality: Philosophical and Computational Foundations*, Oxford: Clarendon Press.

Williamson, J. and Corfield, D. (2001) 'Bayesianism into the 21st Century', in D. Corfield and J. Williamson, eds., *Foundations of Bayesianism*, Kluwer Applied Logic Series, Dordrecht: Kluwer Academic Publishers, pp. 1–16.