MAKING MULTIPLE HYPOTHESES EXPLICIT: AN EXPLICIT STRATEGY FOR COMPUTATIONAL MODELS OF SCIENTIFIC DISCOVERY[1]

Eric G. Freedman

Psychology Department, The University of Michigan-Flint, Flint, MI 48502-2186.
email: freedman_e@msb.flint.umich.edu

Abstract

Generating and evaluating alternative hypotheses is an essential component of scientific discovery. Evidence from cognitive psychology shows that testing multiple hypotheses, unfortunately, is not effective unless subjects state the hypotheses explicitly. The presence of multiple hypotheses facilitates the efficient elimination of incorrect hypotheses because it makes the possibility of disconfirmation more salient. Although several computational models of scientific discovery have included the generation and evaluation of multiple hypotheses, most of these approaches have not focused on multiple hypotheses. Therefore, further research needs to examine systematically the conditions under which multiple hypotheses are effective.

1 INTRODUCTION

Numerous scholars (Chamberlin, 1904; Klahr, Dunbar & Fay, 1990; Platt, 1964) have advocated the evaluation of competing theories in science. Unfortunately, despite this advice, most psychological research (see Klayman & Ha, 1987, for review) has found that scientists and nonscientists have been shown to rely on a positive-test strategy, that is, people generate tests which are an instance of their hypotheses. In most of this previous research, subjects tested only a single hypothesis at a time. Farris and Revlin (1989) have suggested that subjects' reliance on positive tests and their inability to benefit from a disconfirmatory or negative-test strategy (i.e., generating tests that are not an instance of the current active hypothesis) may be attributable to a failure to consider alternative hypotheses. As reviewed below, the use of multiple hypotheses is an effective strategy for scientific induction. Within cognitive science, computational models of scientific discovery have shifted from data-driven models (Langley, Simon, Bradshaw, & Zytkow, 1987) to theory-driven models (Cheng, 1990; Rajamoney, 1990). While some of these theory-driven models have included the generation and evaluation of multiple hypotheses, it will be argued that the evidence from psychological research on the use of multiple hypotheses should be integrated into future computational models of scientific discovery. The present paper will provide some initial recommendations for how the investigation of the role multiple hypotheses can be integrated into the cognitive science of scientific discovery.

2 EVIDENCE FROM COGNITIVE PSYCHOLOGY

Early psychological research which encouraged subjects to consider multiple hypotheses produced mixed results. On the one hand, Klahr and Dunbar (1988, Experiment 2) and Klayman and Ha (1989) have found that asking subjects to consider alternative hypotheses, either before or during hypothesis testing, improved performance. McDonald (1990) found that testing multiple hypotheses was effective only when the target hypothesis was a subset of the subjects' hypotheses. On the other hand, Tweney et al. (1980, Experiment 2) found that encouraging subjects to use multiple hypotheses reduced performance. Subjects had considerable difficulty considering multiple hypotheses and instead considered multiple pieces of evidence against a single hypothesis.

[1]This research was supported in part by a grant/fellowship from the Faculty Research Initiative Program from the University of Michigan-Flint.

Similarly, Freedman (1991a) found that encouraging individual members of groups to consider multiple hypotheses did not improve performance. Instead, groups were successful when they worked together to confirm a single hypothesis at time. One reason for these mixed findings is that subjects may have difficulty maintaining their hypotheses in working memory because they were not required to state their hypotheses explicitly.

To make multiple hypotheses explicit, Freedman (1991b, 1992a, 1992b, 1994; Freedman & Jayaraman, 1993) required that subjects state a pair of hypotheses on each trial. Freedman (1991b) found that multiple hypotheses improved performance when used in conjunction with a negative-test strategy (i.e., generating tests that were not an instance of one's current hypotheses). Freedman (1992b) examined whether the possibility of error in the experimenter's feedback affected the testing of single and multiple hypotheses. He observed that the possibility of error in the experimenter's feedback did not have a detrimental effect when subjects tested multiple hypotheses. Freedman (1992a) also found that when testing multiple hypotheses, interacting four-member groups performed better than individuals in part because groups generated significantly more diagnostic tests. Part of groups' success can be attributed to the fact that they seek and received greater amounts of disconfirmation than individuals. Previous research (Freedman, 1991,1992a) has indicated that the evaluation of multiple hypotheses is facilitated by the generation of diagnostic tests, that is, experiments that are an instance of one hypothesis and are not an instance of the second hypothesis. Testing multiple hypotheses is a more efficient (measured in terms of the number of experiments conducted) method than testing a single hypothesis (Freedman, 1991,1992a, 1992b).

A second line of research has focused on determining what experiment-generation strategies facilitate the use of multiple hypotheses. Freedman and Jayaraman (1993) investigated the evaluation of multiple hypotheses by encouraging either use of a diagnostic strategy (i.e., tests that were an instance of one hypothesis and not an instance of another hypothesis) or generating maximally different hypotheses. Employing a diagnostic strategy improved performance although generating maximally different hypotheses did not. Using a diagnostic strategy decreased positive tests and confirmation, increased negative tests, diagnostic tests, and disconfirmation. Thus, use of a diagnostic strategy facilitated the elimination of incorrect hypotheses.

Freedman (1994) investigated use of a diagnostic strategy (i.e., tests that were an instance of one hypothesis and not an instance of another) during the evaluation of multiple hypotheses in individuals and groups. Interacting groups were more likely to determine the target hypothesis than individuals. Interacting groups employing a diagnostic strategy generated fewer experiments. A diagnostic strategy reduced subjects' reliance on positive tests. In summary, use of multiple hypotheses, in the studies reviewed in this section, increased the efficiency of scientific induction, increased the salience of disconfirmation and decreased reliance of confirmatory tests.

3 COMPUTATIONAL APPROACHES TO THE EVALUATION OF MULTIPLE HYPOTHESES

Several computational models of scientific discovery (Cheng, 1990; Kulkarni & Simon, 1988; Pazzani & Flowers, 1990; Rajamoney, 1990) include the generation of competing hypotheses as part of the discovery process. In most models, a similar process of theory generation and revision is exhibited. For instance, Rajamoney (1990) developed COAST an explanation-based theory revision program. In COAST, theory revision begins with the encountering of an anomalous finding. COAST next proposes revisions to the original theory and then designs experiments to test the proposed theories. New theories must be able to explain the previous findings. In KEKEDA (Kulkarni & Simon, 1988), hypotheses are generated that either include the surprising finding as part of the hypothesis or as a special subprocess. Experiments are designed by generating predictions for each of the proposed theories and selecting an experiment that allows a discrimination among the competing theories. Theories are then eliminated when an observation doesn't satisfy a prediction. Still,

as Cheng (1990) notes, most theory-driven discovery systems tend to make minor modifications in current hypotheses in order to account for the anomalous data. Moreover, computational approaches need to develop different approaches to the evaluation of alternative hypotheses. For instance, Pazzani and Flowers (1990) have suggested that scientific discovery can be conceived as an argumentation process in which alternative theories compete against each other. He has suggested the integration of argument heuristics into computational models.

In most computational models of scientific discovery, hypotheses are eliminated when they can not account for every previous finding. However, it is often the case that scientific theories can not account for every empirical finding. Thagard (1989) has suggested that theories are accepted to the degree that they provide a more coherent explanation of the available evidence than the competing theory. Thagard has developed ECHO, a theory-appraisal program, that employs connectionist algorithms. In ECHO, acceptance of scientific theories is a function of the degree to which a set of hypotheses cohere together to explain the available evidence. Unfortunately, ECHO receives all relevant evidence and hypotheses before it is run. ECHO does not generate experiments or new hypotheses. Still, use of explanatory coherence may provide a worthwhile avenue for the examination of competing hypotheses because most computational models revise current hypotheses by generating new hypotheses that can account for the previous findings. Thus, a systematic examination of various methods for generating and evaluating alternative hypothesis is the next logical stage in the development of theory-driven models of scientific discovery.

4 EXTENSIVE DUAL-SPACE SEARCH MODEL

The results of the psychological and computational research show that making alternative hypotheses explicit has an important heuristic value in scientific induction. Yet, much of the seemingly contradictory findings can be explained because previous researchers have not attempted to develop a model of the cognitive processes underlying the use of multiple hypotheses during scientific induction. Evaluating multiple hypotheses depends on the ability to eliminate incorrect alternative hypotheses. This eliminative ability depends on the search of the experiment and hypothesis space and how subjects modify their hypotheses based on evidence obtained. To illustrate how multiple hypotheses may affect the search of the experiment and the hypothesis spaces, Figure 1 portrays one possible relationship between two proposed hypotheses (H_1 & H_2) and the target hypothesis (T). Consistent with the dual-search space model (Klahr & Dunbar, 1988), Freedman (1992a) has suggested that testing multiple hypotheses leads to a more extensive search of the hypothesis and experiment search spaces. First, asking subjects to consider alternative hypotheses may result in a more extensive search of the hypothesis space. When subjects examine sufficiently distinct alternative hypotheses, they are more successful (Farris & Revlin, 1989; Klahr et al., 1990). Additionally, Freedman (1992a, 1992b) found that subjects who tested multiple hypotheses generated more unique hypotheses than subjects who tested a single hypothesis. Second, a more extensive hypothesis-space search can increase the effectiveness of the experiment-space search by making disconfirmation more salient. When testing a single hypothesis (i.e., H_1), subjects typically generate positive tests in Regions 1, 3, 5, and 6. The presence of multiple hypotheses increased the use of a negative-test strategy and increased the amount of disconfirmation received (Freedman, 1991b, 1992a, 1992b). Because it is impossible to know beforehand whether a test will disconfirm one's hypotheses, the only way to ensure that one of your competing hypotheses is disconfirmed is to conduct a diagnostic test. Consistent with the computational research (Cheng, 1990; Kulkarni & Simon, 1988; Rajamoney, 1990), Freedman (1992a, 1992b, 1994; Freedman & Jayaraman, 1993) found that the evaluation of multiple hypotheses is facilitated by the use of diagnostic tests (i.e., Regions 3, 4, 5, & 7). A diagnostic strategy is useful because regardless of the results one of the current hypotheses will be eliminated. Thus, a

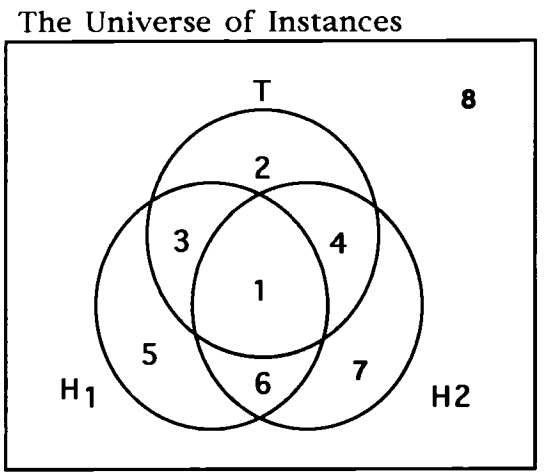

Figure 1. A possible relationship between a target hypothesis (T) and two alternative hypotheses (H_1 & H_2).

diagnostic strategy facilitates the elimination of incorrect hypotheses.

A second crucial factor determining whether testing multiple hypotheses will be effective is how subjects modify their hypotheses in the presence of disconfirmatory evidence. A negative test of one's current hypothesis which is an instance of the target hypothesis indicates that one's current hypothesis is narrower than the target hypothesis (i.e., Region 2). Therefore, theory-revision should produce an increase the breadth of subjects' hypotheses. Conversely, a positive test that is not an instance of the target hypothesis (i.e., Regions 5, 6, & 7) should produce a decrease in the breadth of one's current hypothesis. Thus, the relation between the type of test conducted (i.e., positive vs. negative tests) and the type of feedback received determines where in the hypothesis space one should search for new alternative hypotheses. A third constraint on people's ability to monitor and evaluate multiple hypotheses is their working memory. It is assumed that each hypothesis occupies a certain amount of a limited working memory capacity. In addition, people also need to think how each hypothesis relates to the previous evidence. Because this information occupies working memory, hypothesis testers may have difficulty benefiting from multiple hypotheses. Therefore, individuals with greater working memory should be better able to test multiple hypotheses. Further research is needed to determine whether the ability to test multiple hypotheses is related to working memory capacity.

5 DISCUSSION

The results of psychological research on testing multiple hypotheses support several general conclusions. First, generating multiple hypotheses is not sufficient for facilitating scientific induction, multiple hypotheses need to be made explicit. Second, although testing of multiple hypotheses may not increase the probability of discovering the target hypothesis, subjects often discover the target hypothesis in fewer trials. Thus, use of multiple hypotheses may be a more efficient strategy. Third, multiple hypotheses make the possibility of disconfirmation more salient. Fourth, interacting groups appear to evaluate multiple hypotheses more effectively than individuals. Finally, encouraging use of a diagnostic strategy may be relatively beneficial to facilitation the elimination of incorrect alternative hypotheses. Moreover, the results of the present studies also support the view that multiple hypotheses lead to a more extensive search of the experiment and hypothesis spaces. Thus, this psychological research fits within the framework developed by Newell and Simon. Consequently, the psychological processes employed during the testing of multiple hypotheses need to be integrated into prospective computational models.

While computational models (e.g., COAST) have included the evaluation of competing hypotheses as a component of their models, the results from the present studies suggest that the generation and evaluation of alternative hypotheses should become the explicit focus of future research. The results of the present studies suggest several explicit strategies for further work in computational models of scientific discovery. First, various experiment-space search strategies should be compared. Even though both humans and computational models benefit from a diagnostic strategy, the breadth of the experiment space searched should be manipulated by varying the number of features that a new experiment has

in common with previous experiments. Second, the way in which the hypothesis space is search should be investigated systematically. Specifically, various hypothesis-generation heuristics should be developed and compared. On the one hand, when faced with an anomaly or disconfirmatory finding, most computational models generate new hypotheses that reflect minor changes in the current hypothesis (cf. Cheng, 1990). From the perspective of Klahr et al., most computational models (e.g., COAST, KEKEDA) generate new hypotheses from within the same region of the hypothesis space. On the other hand, Klahr et al. has found that human subjects who consider hypotheses from different parts of the hypothesis space were more likely to discover the target hypothesis compared to those who made minor hypothesis changes. Although STERN (Cheng, 1990) does evaluate mutually exclusive hypotheses, computational models of scientific discovery should manipulate the scope of the hypotheses generated. Hypothesis-generation heuristics could be developed that vary the degree of theory revision. A theory-revision heuristic could be developed that generates new hypotheses that are maximally dissimilar to current hypotheses. For instance, further research should compare models that either generate new hypotheses that include only the anomalous finding or models that include the anomalous finding plus many other instances. Furthermore, new hypotheses should be selected from different regions of the hypothesis space than the current hypotheses. Besides accounting for the previous findings, the generation of new hypotheses should be guided by the attempt to increase the explanatory coherence. STERN does include explanatory breadth as one of its criterion for the evaluation of hypotheses. Third, the influence of working memory on evaluating multiple hypotheses should be examined. Clearly, like humans, computational models should be limited in the number of hypotheses that can be evaluated at any time. It is likely that the performance of a theory-driven computational model of scientific discovery would be affected by the amount of working memory capacity available. Fourth, the research reviewed above indicates that interacting groups are superior to individuals working alone. The present research shows that the framework developed by Newell and Simon can be extended to a group context. Therefore, as I've (Freedman, 1992c) previously argued, computational models of science need to include social factors in their models. Group problem solving could be integrated into computational models of scientific discovery by distributing the discovery process among several communicating programs. Finally, various hypothesis-evaluation heuristics should be compared. Besides heuristics that eliminate or modify theories when the evidence is inconsistent with its predictions, the explanatory coherence, simplicity and breadth of a theory should be considered.

The implications of this research for computational approaches to science is evident. Given the complexity of scientific theories and scientists' information-processing limitations, computational models may serve as an important adjunct to researchers. Intelligent-based systems should be utilized to facilitate the generation and elaboration of alternative representations. Once articulated, experiment-generation systems (e.g., KEKEDA) could design experiments that could help to discriminate among competing hypotheses. Again, given the relative efficiency of a multiple-hypothesis strategy, the potential benefits to science and cognitive science could be enormous.

REFERENCES

Chamberlin, T. C. (1904). The method of earth science. Popular Science Monthly, 66, 66-75.

Cheng, P. C-H. (1990). Modeling scientific discovery. Technical Report No. 65, Human Cognition Research Laboratory, The Open University, Milton Keynes, England.

Farris, H., & Revlin, R. (1989). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. Memory and Cognition, 17, 221-232.

Freedman, E. G. (1991a). Role of multiple hypotheses during collective induction. Paper presented at the 61st Annual Meeting of the Eastern Psychological Association, New York, NY.

Freedman, E. G. (1991b). Scientific induction: Multiple hypotheses and test strategy. Paper presented at the 32nd Annual Meeting of the Psychonomic Society, San Francisco, CA.

Freedman, E. G. (1992a). Scientific induction: Individual versus group processes and multiple hypotheses. Proceedings of the 14th Annual Conference of the Cognitive Science Society, 183-188, Hillsdale, NJ: Lawrence Erlbaum.

Freedman, E. G. (November, 1992b). The effects of possible error and multiple hypotheses on scientific induction. Paper presented at the 33rd Annual Meeting of the Psychonomic Society. St. Louis, MO.

Freedman, E. G. (1992c). Understanding scientific controversies from a computational perspective: The case of latent learning. In R. Giere (Ed.), Cognitive Models of Science: Minnesota Studies in the Philosophy of Science, (Volume XV) (pp. 310-337). Minneapolis, MN: University of Minnesota Press.

Freedman, E. G. (November, 1994). Use of a diagnostic strategy in individuals and interacting groups. Paper accepted for presentation at the 35th Annual Meeting of the Psychonomic Society. St. Louis, MO.

Freedman, E. G., & Jayaraman, R. (November, 1993). Evaluating multiple hypotheses with a diagnostic strategy and maximally different hypotheses. Paper presented at the 34th Annual Meeting of the Psychonomic Society. Washington, D.C.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-48.

Klahr, D., Dunbar, K., & Fay, A. L. (1990). Designing good experiments to test bad hypotheses. In J. Shrager & P. Langley (Eds.), Computational models of scientific discovery and theory formation (pg. 356-402). Palo Alto: Morgan Kaufmann.

Klayman, J., & Ha, W. (1987). Confirmation, disconfirmation and information in hypothesis testing. Psychological Review, 94, 211-228.

Klayman, J., & Ha, Y-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 317-330.

Kulkarni, D., & Simon, H. A. (1988). The process of scientific discovery: The strategy of experimentation. Cognitive Science, 12, 139-175.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). Scientific Discovery: Computational Explorations of the Creative Process. Boston, MA: MIT Press.

McDonald, J. (1990). Some situational determinants of hypothesis-testing strategies. Journal of Experimental Social Psychology, 26, 255-274.

Pazzani, M. J., & Flowers, M. (1990). Scientific discovery in the layperson. In J. Shrager & P. Langley.(Eds.), Computational models of scientific discovery and theory formation (pg. 403-435). Palo Alto: Morgan Kaufmann.

Platt, J. R. (1964). Strong Inference. Science, 146, 347-353.

Rajamoney, S. (1990). A computational approach to theory revision. In J. Shrager & P. Langley.(Eds.), Computational models of scientific discovery and theory formation (pg. 225-253). Palo Alto: Morgan Kaufmann.

Thagard, P. (1989). Explanatory coherence. Behavioral and Brain Sciences, 12, 435-502.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A. & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. Quarterly Journal of Experimental Psychology, 32, 109-123.