Order Effects in Incremental Learning

1. Introduction

Intelligent agents, including humans, exist in an environment that changes over time. Thus, it seems natural that models of learning in such agents take into account the fact that this process also takes place over time. We often refer to such agents as *incremental* learners, in that the temporal nature of experience leads them to incorporate that experience in a piecemeal fashion.

In this chapter we discuss the notion of incremental learning from three perspectives – machine learning, instructional theory, and experimental cognitive psychology. These fields share a concern with the incremental nature of learning and with the effects of training order on the acquired knowledge. However, the literature has often been imprecise and sometimes inconsistent about the definition and nature of incremental learning, suggesting the need for a clearer treatment of the issues that arise in this context.

We attempt to clarify the situation in the following section by presenting some definitions of incremental learning and introducing some distinctions among types of order effects. We then turn to a more detailed discussion of two such types of effects from the vantage points of the different fields, briefly reviewing some relevant work in each case. Finally, we outline some directions for future research on this intriguing topic.

2. The Nature of Incremental Learning

Most research on incremental learning rests on three assumptions, which are often implicit in the literature in this area. Each of these assumptions appears to hold for human learners, and they seem equally desirable for artificial ones. First, the agent should be able to use its learned knowledge to carry out its performance task at any stage of learning. Second, the incorporation of experience into memory during learning should be computationally efficient. Finally, the learning process should not make unreasonable space demands, so that memory requirements increase as a tractable function of experience.[1]

2.1 Definitions of Incremental Learning

The literature also contains different definitions of incremental learning, sometimes implicit, which seem tied to the above assumptions. We should briefly review these alternatives, in hopes of deciding which one is most appropriate for our current analysis. Perhaps the most common definition deals only with the first of the above assumptions.

[1] In most of examples, the term "experience" translates to "training instance". But because we will see other senses of the former term elsewhere in the chapter, we will use it in our definitions.

Definition 1 *A learner L is* incremental *if L inputs one training experience at a time.*

Clearly, for any learner of this sort, one can interrupt the training process and ask the agent to use its acquired knowledge to make predictions or carry out some other task. Such a learner certainly appears incremental to an external viewer.

However, note that one can easily adapt *any* learning algorithm to fit this definition, including ones that process many instances at a time, by simply storing the instances observed so far and running the method on them. For example, Schlimmer and Fisher (1986) describe such a variant of Quinlan's (1986) nonincremental ID3 algorithm for decision-tree induction. This system simply runs ID3 as a subroutine on the first training case, the first two cases, the first three cases, and so on, thus mimicking the external behavior of an incremental learner. One can adapt this idea to make any nonincremental learning algorithm appear incremental by our first definition. In fact, the above definition says more about the nature of the learning task than about the learner itself. Within the machine learning literature, particularly that on computational learning theory, this situation is sometimes referred to as an *online* learning problem (Littlestone, 1987).

Clearly, it seems desirable to distinguish between arbitrary methods that handle online tasks and ones that better reflect our intuitions about incremental processing. A more plausible definition would incorporate not only the first assumption but also the second one given above.

Definition 2 *A learner L is* incremental *if L inputs one training experience at a time and does not reprocess any previous experiences.*

This version actually places a constraint on the learning mechanism itself, in that it can process each experience only once. We might relax this constraint somewhat to allow limited reprocessing, provided we do so cautiously. The important idea is that the time taken to process each experience must remain constant or nearly so with increasing numbers, in order to guarantee efficient learning of the sort seen in humans.[2]

Although this definition is a considerable improvement, it still violates some important intuitions. For example, Mitchell's (1982) candidate elimination algorithm for learning logical conjunctions processes instances one at a time and does not need to reprocess them. However, it accomplishes this feat by retaining in memory a set of competing hypotheses that summarize the data, and it reprocesses these hypotheses upon incorporating each training case. This presents no problem by itself, but Haussler (1987) has shown that the number of such hypotheses can grow exponentially with the number of training items, which seems contrary to our notions of incrementality.

[2] Note that even when learning method is incremental in this sense, one may not use it in an online fashion. For example, the weight-updating scheme used in backpropagation for neural networks does not reprocess instances by itself, yet researchers typically rerun the algorithm over the training set many times, thus violating our second assumption about reprocessing.

We can avoid the inclusion of such algorithms by incorporating the third of the above assumptions into our definition.

Definition 3 *A learner L is an* incremental *if L inputs one training experience at a time, does not reprocess any previous experiences, and retains only one knowledge structure in memory.*

This formulation rules out learning methods that retain competing descriptions, such as the candidate elimination algorithm, as well as methods like Winston's (1975) that carry out explicit backtracking. Learners that are incremental in this sense retain no set of alternatives and no memory of where they have been; they can only incorporate the next training item and move forward in response. For this reason, Langley, Gennari, and Iba (1987) refer to them as *incremental hill climbing* approaches to learning.

We will restrict ourselves to this third definition of incremental processing in the remainder of this paper. We maintain that any viable theory of human learning must be based on this definition, and we will see that many common learning methods satisfy it, though they are seldom presented in these terms. We can loosen our definition somewhat to allow storage of a few competing knowledge structures, or to allow a current structure with a number of possible successors, from which one is then selected. These variations still restrict memory to a manageable size.

2.2 Definitions of Order Effects

Learning mechanisms that rely on incremental hill climbing have one central characteristic that has received considerable attention: their behavior tends to be affected by the *order* of experience. We can state this notion more precisely:

Definition 4 *A learner L exhibits an* order effect *on a training set of experiences T if there exist two or more orders of T for which L produces different knowledge structures.*

The origin of such effects is best understood in terms of search through the space of knowledge structures. An incremental learning method must make decisions about which path to follow (which structure to create) based on samples of the data. Different early samples may lead the learner down quite different paths, and later experiences may not be sufficient to counteract them.[3] Figure 1 (a) shows the paths through the space of knowledge structures for two different orders of the same experiences; because the learner arrives at different structures, this constitutes an example of an order effect. In contrast, the paths in Figure 1 (b) diverge initially but lead to the same structure, meaning no order effect has occurred.

[3] We must distinguish between behavior differences that result from order effects on a given training set and the quite distinct ones that result from different samples of data. The latter can occur even with the most nonincremental of learning methods.

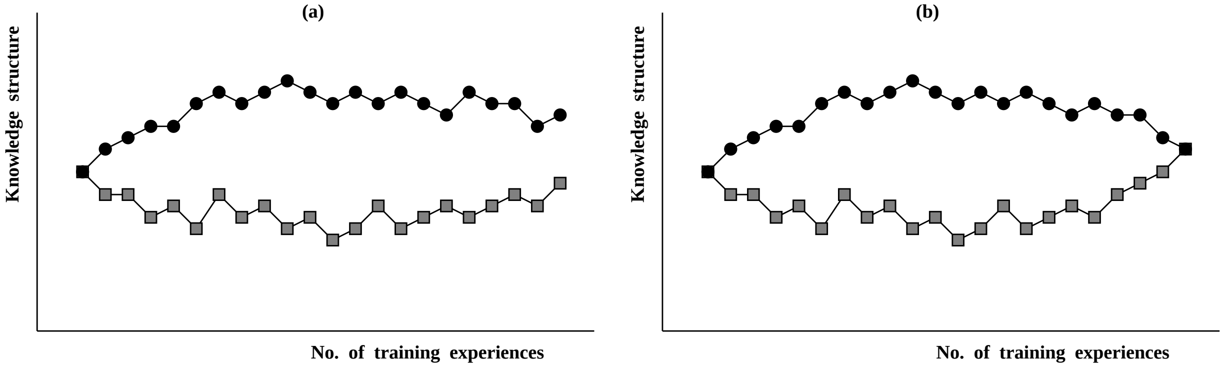

Figure 1. The knowledge structures generated by an incremental learner for two different orders of training experiences when (a) an order effect produces two different structures and (b) when the absence of an order effect produces the same structure.

Our definition of order effects focuses on particular training sets and their presentation order, but we can rephrase things to emphasize the algorithm itself:

Definition 5 *An learner L is* order sensitive *if there exists a training set T on which L exhibits an order effect.*

Similarly, we can say that a learner is *order independent* if it never exhibits an order effect. This formulation takes an all-or-none stance, but clearly one can also talk about degrees of order sensitivity, in terms of the number of training sets and the number of orders in which such effects occur, as well as the resulting distance between the learned structures.

One can also talk about the implications of order effects on behavior, using some performance measure M that reflects the usefulness of the knowledge learned from experience, such as accuracy on test cases. It seems reasonable to assume that some order effects, although producing different knowledge structures, have relatively little impact on performance.

Definition 6 *An order effect for learner L on training set T is* benign *with respect to measure M if all orders of T produce knowledge structures of (nearly) equal scores on M.*

In contrast, we can say that an order effect is *malignant* if different orders produce quite different results on the performance measure. Naturally, malignant order effects hold greater interest for most learning researchers, especially those with prescriptive rather than descriptive goals.

2.3 Levels of Order Effects

There exist at least three different levels at which order effects can occur, and thus three different ways in which we can instantiate the term *experience* in the previous definitions. Recall that most

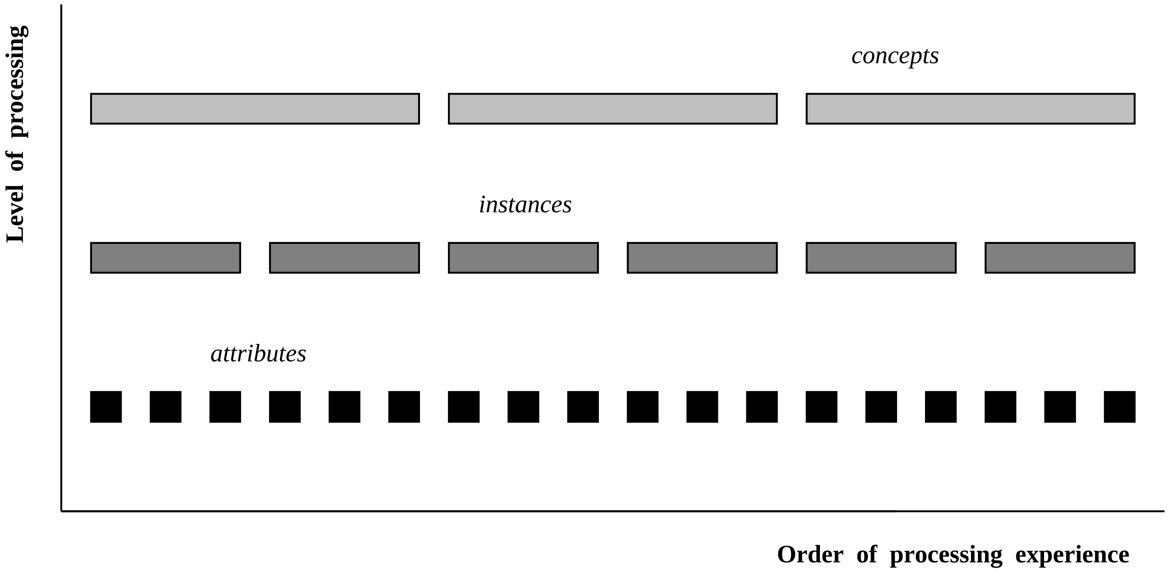

Figure 2. Incremental processing can occur at three levels of temporal resolution: with respect to the attributes used to describe to instances, for instances of the concepts being learned, and for the concepts themselves.

learning research deals with the acquisition of *concepts* from training *instances* that are described in terms of *attributes* or *features*. Incremental processing can occur, and thus order effects can result, with respect any of these levels, as depicted in Figure 2.

At the finest temporal resolution, the agent can process the attributes of each instance one at a time. In some frameworks, such as discrimination networks, attribute order can affect both performance and learning. Clearly, humans have limited attentional resources, so that one might expect that researchers would give high priority to modeling the effect of attribute order. Nevertheless, though many systems give different importance to different attributes, only a few (e.g., Feigenbaum, 1963; Gennari, 1991) acknowledge that they can be observed in different orders, and even these do not explicitly examine the effects of observation order on learning. For this reason, we will not have much to say on the topic here.

At the intermediate level, the agent can process training instances one at a time. This is clearly the most common interpretation of both incremental learning and order effects within the literature, and we consider it at some length in the next section. We will see that there exist machine learning algorithms that process instances in an incremental manner, psychological studies of the effects of instance order, and hypotheses about the uses of instance order in education, each of which sheds a different light on the nature of incremental processing.

At the highest level, the agent can learn distinct concepts one at a time, and their order of acquisition can make the learning task more or less difficult. There exists some work on this topic within machine learning and cognitive psychology, but it has received perhaps the most attention within the education paradigm, where courses of instruction typically order concepts in some principled fashion. We devote Section 3 to the incremental learning of different concepts.

3. The Effects of Instance Order

Research on the effects of instance order can approach the problem from different perspectives. Much of the work takes a prescriptive slant, treating order effects either as something to be eliminated or something to use profitably. Another alternative is to treat order effects as a phenomenon to be studied from a purely descriptive angle. Below we consider each of these vantages in turn.

3.1 Mitigating the Effects of Instance Order

If one's aim is to engineer an autonomous agent that learns from experience in a robust manner, then the effects of instance order – at least malignant ones – are undesirable. For this reason, many machine learning papers on incremental methods discuss schemes for eliminating or mitigating the order sensitivity of induction algorithms. Researchers have explored a variety of approaches to this issue within the framework of incremental hill climbing that we defined in Section 2.

The simplest scheme involves making strong assumptions about the nature of the target concept, so that different orders of the same training data always produce the same result. For example, some early work on the induction of logical concepts focused on *conjunctive* concepts, and algorithms for this task which move from specific to general hypotheses show no order sensitivity, at least when used on attribute-value descriptions. Similarly, the naive Bayesian classifier (Langley, Iba, & Thompson, 1992) assumes both a single probabilistic summary for each class and independence among attributes; this lets it use a simple learning method that updates counts for each observed combination of class and attribute value, which makes it completely insensitive to training order. Within the area of grammar induction, Angluin (1977) describes an incremental algorithm for learning the restricted class of k-reversible grammars; this technique adds a new chain of states to an existing finite-state machine for each sentence it encounters, then merges states in a way that also guarantees against order effects.

However, some researchers find such representational restrictions distasteful (despite their excellent performance on many domains), and so have considered other responses to the problem. An alternative approach relies on background knowledge to constrain the learning process and thus to mitigate order effects.[4] When used to improve classification accuracy, explanation-based methods constitute an extreme version of this idea (e.g., Flann & Dietterich, 1989). In this scheme, each training case leads to the creation of one rule, and the order in which they are added to memory

[4] Of course, because incremental methods process experiences sequentially, the results of learning from the first n instances constitutes a form of background knowledge for the $n + 1$st instance. Here we refer to knowledge that is available before the learning process begins.

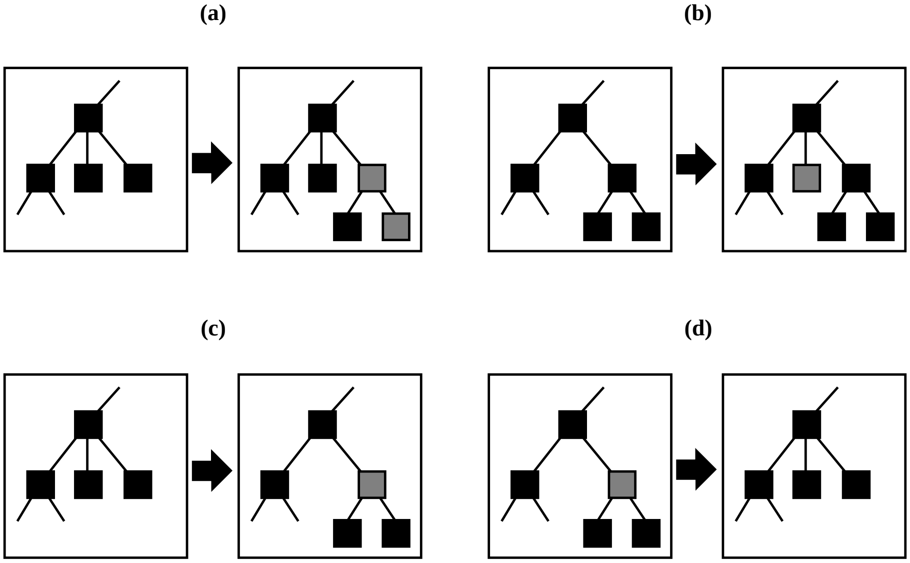

Figure 3. Learning operators used by Fisher's (1987) Cobweb to modify the structure of a probabilistic concept hierarchy: (a) extending the hierarchy downward; (b) creating a new sibling at the current level; (c) merging two existing concept; and (d) splitting an existing concept. Operators (c) and (d) are explicitly designed to reduce order effects.

does not affect the result. Less extreme variations of this approach are also possible. For example, McKusick and Langley (1991) show that providing a partial concept hierarchy can reduce order effects in incremental clustering systems, and we expect similar results would occur with other learning methods. Cornuejols (1993) presents an insightful formal analysis of the conditions under which background knowledge reduces order effects.

Yet another response incorporates bidirectional learning operators that can undo early decisions that were based on nonrepresentative data. For example, many logical induction methods include both operators for making rules more general and more specific (e.g., Iba, Wogulis, & Langley, 1988), and algorithms for hierarchical clustering often include both operators for merging and splitting nodes in a taxonomy, as shown for Fisher's (1987) Cobweb algorithm in Figure 3. Such dual operators give learners the ability to emulate backtracking in certain situations, even though they have no memory of their previous knowledge structures. Gennari et al. (1989) present evidence that such operators reduce order effects in one clustering system, and their inclusion in other systems suggests their usefulness there as well.

A fourth framework takes a more conservative approach, retaining a current hypothesized knowledge structure and a manageable set of potential successor hypotheses, then waiting until one has observed enough training cases to determine with confidence the best successor. For example, Schlimmer and Fisher's (1986) approach to incremental decision-tree induction retains statistics on alternative attributes and extends the tree downward only when one attribute appears statistically better than the others. In a similar manner, Iba (1987) collects statistics on macro-operators, making sure they will aid problem-solving efficiency before adding them to memory. Greiner (1992) describes a very general formulation of this idea; his approach to incremental hill climbing needs no bidirectional operators because it collects enough training cases to (nearly) always makes the right decision at each step in the search process.

Another class of methods attempts to reduce order effects by storing multiple, complementary descriptions in memory. Like the previous approach, this relies on the idea that one should not commit to a single knowledge structure early on, before the training data clearly indicate that one is better than the others. For example, some clustering systems create directed acyclic graphs that can sort instances in multiple ways (e.g., Levinson, 1985; Martin & Billman, 1994). In its extreme form, this approach violates the spirit of incremental hill climbing in that the size of the knowledge structure can grow very rapidly as a function of the number of training cases. However, placing restrictions on the number of complementary descriptions can ensure manageable memory while retaining the benefits of a least-commitment approach.

A final approach concerns the notion of distributed representations in multilayer neural networks. The standard method for learning in such networks, backpropagation, is typically run through the training set many times, rather than altering the network to master each case before presenting the next one. McClosky and Cohen (1989) and Ratcliff (1990) have shown that, when trained in this latter mode, backpropagation exhibits "catastrophic interference", in that learning each new item causes the network to forget those learned previously. French (1993) argues that this effect results from the distributed representation of knowledge embodied in the network's hidden units. He shows that one can reduce this order effect by encouraging backpropagation to produce weights that give less distributed activations among hidden units. French also suggests that Kruschke's (1993) approach, which uses an inverse exponential activation function, avoids catastrophic interference for the same reason.

3.2 Instance Order in Instructional Design

Not all researchers view order effects as something to eliminate; some instead see them as given constraints that one must take into account during instruction. This attitude is especially prevalent among those in education who study instructional design, but it also occurs in some work on machine learning. Research in this tradition involves the design of presentation orders that will maximize the rate or quality of learning by simplifying the acquisition task.[5]

One simple technique along these lines involves the idea of a *near miss*, which is a negative instance of some concept that almost but does not quite satisfy the concept's definition. Winston (1975) posits that presenting a positive training case followed by near misses will simplify the induction process, as this lets the learner easily detect individual differences between the positive

[5] One should not confuse the positive use of order effects during instruction with the notion of benign effects discussed above. In fact, malignant effects are most relevant for instructional design.

instance and the negative ones that are necessary to the concept definition. Unfortunately, this technique makes sense only for concepts with logical definitions, making it inapplicable to many natural categories and to many learning methods. Neri and Saitta (1993) describe a more general technique for selecting training cases to increase learning rates.

Another variant on this approach assumes that learners will fare better if presentation order alternates among instances of different categories than if presented with many cases of one category followed by those of another. The intuition here is that observation of contrasting training cases will encourage introduction of the appropriate distinctions. McKusick and Langley (1991) show that this training regimen increases the learning rate for a probabilistic clustering method by encouraging the creation of proper distinctions high in the concept hierarchy. A related idea underlies the standard 'epoch' training for backpropagation in neural networks, which iterates through training cases many times in order to avoid the "catastrophic interference" described earlier.

VanLehn (1987) observes that, in teaching complex concepts and procedures, many textbooks present them "one disjunct at a time". That is, if the target structure involves a number of distinct components that cover different situations, these components are presented separately, with each one being learned before turning to the next. The assumption here is that learning a number of simple concepts or procedures, when identified as such, is easier than acquiring a single complex structure. Superficially, at least, this advice seems to conflict with the alternation strategy recommended above, though VanLehn focused on procedural tasks and the alternation scheme comes from work on concept learning.

In some domains, training cases can themselves have different levels of complexity. Porat and Feldman (1991) take advantage of this fact to simplify the problem of grammar induction by presenting simple sample sentences before more complex ones. Like the "one disjunct" strategy, this training order lets the learner master parts of the target grammar on simple cases before being challenged by harder ones. Elman (1991) uses a similar training regimen for connectionist learning of phrase structure grammars, in which he gradually increases the proportion of complex sentences in the training set. This approach also seems recommended for memory-limited agents (including humans), as it lets them establish chunks during the early phases that aid retention during later periods.

Rendell (1986) and Iba (1989) draw on a similar insight in their work on learning problem-solving strategies. Their systems are initially presented with relatively easy problems which they can solve with little domain knowledge; they then use the resulting solution as material for learning. Once the systems have acquired some knowledge in this manner, they are given more difficult problems to drive further learning. This "bootstrapping" approach would seem generally useful whenever the agent must learn from the results of some search process.

Another approach to instruction hypothesizes an *accretion* theory of learning in humans (Rumelhart & Norman, 1978), in which new knowledge is added to existing structures. According to this theory, new experiences that have some connection with known memory structures are stored and accessible, whereas experience that makes little or no contact is effectively lost. This suggests a training regimen in which one first presents information about core ideas, then gradually presents elaborations that build the learner's knowledge base outward around the edges.

Finally, some research in both education and machine learning emphasizes the power of letting the learner select its own experiences. Plotzner (1990) presents evidence that students acquire knowledge about physics more rapidly when they control the order of presentation. Similarly, Carbonell and Gil (1987), Gross (1991), and Scott and Markovitch (1991) show the advantages of experimental control for automated learning systems. The basic insight behind this approach is that the learner often knows more about its own hypotheses, and thus about the areas of uncertainty, than does the instructor. Thus, learner-selected instances can provide more information and thus faster acquisition than teacher-selected experience.

3.3 Instance Order and Human Behavior

Until now, we have focused on techniques for reducing order effects in machine learning systems and intuitively plausible approaches for taking advantage of such effects during instruction. Both of these perspectives have a prescriptive flavor. However, there also exist experimental results about the effect of instance order on human learning, though descriptive studies of this sort have been relatively rare.

The literature on human memory touches on order effects, though not as directly as one might like. There is clear evidence of both retroactive inhibition (e.g., Müller & Pilzecker, 1900) – that learning on later items can hurt retrieval of ones mastered earlier – and the related phenomenon of proactive inhibition – that items mastered early on can inhibit learning on later ones. Both forms of interference are more likely when the items are similar in some fashion, thus allowing confusion. Studies of mass vs. distributed practice are also somewhat relevant to order effects, though they were not designed with this issue in mind. Most important, studies of this sort have emphasized rote memorization rather than concept acquisition or similar forms of induction.

A few studies of category learning have dealt with order effects more explicitly. Elio and Anderson (1984) considered two presentation orders of training instances for categories with graded structure, in which some cases were more typical than others. In one condition, they first presented a sample of highly typical instances, followed by a sample containing a mixture of typical and less typical cases, and finally a sample that was fully representative of the category. In another condition, each successive training sample was representative. The authors found that, when instructed to formulate explicit hypotheses, subjects learned the target concepts better (in terms of accuracy and typicality ratings) from purely representative samples. In contrast, when told to simply remember the individual instances, they did better when first seeing only typical samples. In later work, Elio and Lin (1994) modeled this interaction effect with two distinct learning strategies, one involving rule induction and the other using an instance-based mechanism.

Clapper and Bower (1994) report interesting results on an unsupervised learning task involving two simple logical conjunctions, using a performance criterion measuring ability to distinguish attributes with constant values for each category from those which vary. They found that subjects given training cases from one category followed by those from the other category learned more rapidly than did subjects given training instances that interleaved the two categories. Clapper and Bower suggested that the first presentation order lets the learner acquire norms for one category, and then be surprised when instances from the second category depart from those norms. Note that this finding directly contradicts the alternation method discussed earlier, which was found to aid some machine learning algorithms; thus, it suggests that these methods provide poor models of human learning, at least in this domain.

Studies of human problem solving have also revealed some intriguing effects of problem order. Luchins' (1942) experiments with the water jug task showed that, when given a set of training problems that had only one solution, subjects later solved other problems that had alternative solutions in the same way. However, this *Einstellung* effect did not occur when they encountered both types of problems early in training. The order of problem presentation also influenced the time it took subjects to extinguish this behavior. Jones (1989), Langley and Allen (1991), and others provide computational accounts of Luchins' basic phenomena in terms of early acquisition of search-control knowledge and subsequent use of that knowledge to bias problem solving.

In summary, experimental psychology has given less attention to the effects of instance order than machine learning or educational theory, but the studies that have been reported call into question some of the assumptions of the latter two fields. Clearly, a fuller descriptive account of the incremental nature of human learning would complement, and possibly redirect, the prescriptive work in other areas.

4. The Effects of Concept Order

As we noted earlier, one can impose an ordering not only the training cases provided for learning, but also on the concepts that are to be learned. This issue arises only in more complex learning tasks, where some concepts can be defined or grounded in terms of other concepts.[6] Work in this tradition has focused almost exclusively on the advantages of certain presentation orders, rather than on the effects of different orders or on techniques for overcoming them.

4.1 From Simple to Complex Concepts

Most research in this area has assumed that the preferable training order moves from simple to complex concepts. That is, if some high-level concept can be formulated in terms of lower-level

[6] We intend the term "concept" here in the broadest sense possible, to cover not only static structures but also temporal ones like procedures and grammars.

ones, then learning will be aided if one masters the simpler concepts first. A number of machine learning efforts build on this idea. For example, Sammut and Banerji (1986) describe an incremental approach that first learns simple logical concepts from supervised data, then uses the learned rules to reexpress instances of more complex concepts at higher levels of abstraction. Elio and Watanabe (1991) report another learning algorithm that operates along similar lines. In both cases, the intuition is that augmenting the instance description with new features, which must first be learned themselves, can ease the task of learning complex concepts.

The educational theorist Gagne (1966) proposed that human learning occurs in a similar manner, and recommended the design of instructional sequences in which students mastered component skills before attempting to learn about the more complex procedures that require them.[7] For the domain of solving first-order algebraic equations, he presented a skill hierarchy with eight distinct levels, ranging from symbol recognition and number use at the lowest tier, through intermediate skills like simplifying functional expressions and adding numbers to both sides, to equation solving at the highest level. However, as Singley and Anderson (1989) note, experimental support for Gagne's theory has generally been difficult to obtain. In particular, some "scramble" studies have failed to find any differences between training regimens that incorporate the component-first scheme and ones that order skills randomly.

4.2 From Complex to Simple Concepts

Not all researchers have assumed that components are best learned before composite concepts or skills. Shapiro's (1987) technique of *structured induction* recommends exactly the opposite, at least in the use of decision-tree induction to construct knowledge bases. He found that domain experts could provide examples for use as training cases, but that they preferred to describe them in terms of high-level attributes. After constructing a decision tree from these cases, one could then get experts to provide lower-level examples as training data for concepts that defined the initial attributes. The result of this top-down process is a recursively defined decision tree that eventually grounds out in observable features. Langley and Simon (in press) note that structured induction has been used to construct a number of fielded knowledge bases.

Some educational psychologists have proposed analogous training sequences for human instruction. For example, both Bruner's (1966) spiral curriculum and Reigeluth and Stein's (1983) elaboration theory recommend that students first be taught very general skills, and only them be shown techniques for instantiating them. There have been fewer experimental evaluations of such top-down organizations than for bottom-up ones like Gagne's, but clearly they deserve equal attention, as do mixed instructional strategies.

[7] Some theories of human learning, such as Rosenbloom and Newell's (1987) chunking account, assume that simple concepts are learned before more complex ones but take no position on the effect of training order.

5. Directions for Research on Incremental Learning

The study of incremental learning and order effects clearly has important implications for both the construction of artificial intelligent agents and the design of instructional sequences for humans. The work to date has revealed some promising approaches in both areas, but much more remains to be done before we understand the full nature of incremental learning. We can group the important lines of research still needed into three main classes.

First, the field needs better measures for detecting order effects in incremental learners, whether human or machine. Learning curves, which describe performance as a function of the number of training experiences, will likely occupy a central role in this effort. Typically, the mean values of learning curves are used to reveal the rate of learning, but such curves can also suggest the presence of order effects. Briefly, one can expose learners to the same training data in different orders, then examine not the mean but the variance along the resulting curve; a high variance suggests a strong sensitivity to training order in the learner. However, this is only one promising technique, and others may prove just as useful.

We also need better descriptive languages for characterizing the paths taken by incremental learners through the space of knowledge structures, and better techniques for identifying the choices responsible for order effects. For this, we need to look more closely at individual training orders and to compare the behaviors they produce in the learner. If order effects are absent within certain subsets set of training orders but present across these subsets, then the differences among those sets may be useful in describing the cause of the effects. Again, this approach is only one among many possible methods for analyzing sequential behavior in learning.

Finally, we need better theories about the sources of order sensitivity that hold across broad classes of incremental learners. Existing accounts revolve around notions of alternating vs. batch orders, typical vs. atypical instances, and simple vs. complex problems. These provide reasonable starting points, but they are more like simple hypotheses than coherent theories. The view of incremental learning as hill-climbing search through a space of knowledge structures, with decisions affected by the most recent experience, holds the most promise for a unified account of order effects, though the exact nature of this account remains far from clear.

We encourage researchers from education, cognitive psychology, and machine learning to look more closely at the nature of incremental processing, and to build on the growing body of work in this area. We hope that a joint effort by scientists from all three disciplines will lead to insights that would not be possible by studying the effects of training order from a single perspective.

Acknowledgements

Thanks to members of the ESF Task Force on Sequencing Effects for enlightening discussions that led to many of the ideas presented in this chapter. This work was supported in part by a grant from the European Science Foundation on Learning in Humans and Machines, and in part by Grant F49620-94-1-0118 from the Computer Science Division of the U.S. Air Force Office of Scientific Research.

References

Angluin, D. (1977). Inference of reversible grammars. *Journal of the Association for Computing Machinery 29*, 741–765.

Bruner, J. (1966). *Toward a theory of instruction.* New York: W. W. Norton.

Carbonell, J. G., & Gil, Y. (1987). Learning by experimentation. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 256–266). Irvine, CA: Morgan Kaufmann.

Clapper, J. P., & Bower, G. H. (1994). Category invention and unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 443–460.

Cornuejols, A. (1993). Getting order independence in incremental learning. *Proceedings of the 1993 European Conference on Machine Learning* (pp. 196–212). Vienna: Springer-Verlag.

Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & Cognition*, *12*, 20–30.

Elio, R., & Lin, K. (1994). Simulation models of the influence of learning mode and training variance on category learning. *Cognitive Science*, *18*, 185–219.

Elio, R., & Watanabe, L. (1991). An incremental deductive strategy for controlling constructive induction in learning from examples. *Machine Learning*, *7*, 7–44.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139–172.

Flann, N. S., & Dietterich, T. G. (1989). A study of explanation-based methods for inductive learning. *Machine Learning*, *4*, 187–226.

French, R. M. (1993). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. *Proceedings of the AAAI Spring Symposium on Training Issues in Incremental Learning* (pp. 70–77). Stanford, CA: AAAI Press.

Gagne, R. M. (1966). *The conditions of learning*. New York: Holt, Reinhart, & Winston.

Gennari, J. H. (1991). Concept formation and attention. *Proceedings of the Thirteenth Conference of the Cognitive Science Society* (pp. 724–728). Chicago: Lawrence Erlbaum.

Gennari, J. H., Langley, P., & Fisher, D. H. (1989). Models of incremental concept formation. *Artificial Intelligence*, *40*, 11–61.

Greiner, R. (1992). Probabilistic hill-climbing: Theory and applications. *Proceedings of the Ninth Canadian Conference on Artificial Intelligence* (pp. 60–67). Vancouver: Morgan Kaufmann.

Gross, K. P. (1991). *Concept acquisition through attribute evolution and experimental selection*. Doctoral dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Haussler, D. (1987). Bias, version spaces, and Valiant's learning framework. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 324–336). Irvine, CA: Morgan Kaufmann.

Iba, G. A. (1989). A heuristic approach to the discovery of macro-operators. *Machine Learning*, *3*, 285–317.

Iba, W., Wogulis, J., & Langley, P. (1988). Trading off simplicity and coverage in incremental concept learning. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 73–79). Ann Arbor, MI: Morgan Kaufmann.

Jones, R. (1989). *A model of retrieval in problem solving*. Doctoral dissertation, Department of Information & Computer Science, University of California, Irvine.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Langley, P., & Allen, J. A. (1991). Learning, memory, and search in planning. *Proceedings of the Thirteenth Conference of the Cognitive Science Society* (pp. 364–369). Chicago: Lawrence Erlbaum.

Langley, P., Gennari, J. H., & Iba, W. (1987). Hill-climbing theories of learning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 312–323). Irvine, CA: Morgan Kaufmann.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 223–228). San Jose, CA: AAAI.

Langley, P., & Simon, H. A. (in press). Applications of machine learning and rule induction. *Communications of the ACM*.

Levinson, R. A. (1985). *A self-organizing retrieval system for graphs*. Doctoral dissertation, Department of Computer Sciences, University of Texas, Austin.

Littlestone, N. (1987). Learning quickly when irrelevant attributes abound: A new linear threshold algorithm. *Machine Learning*, *2*, 285–318.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, *54* (248).

Martin, J. D., & Billman, D. O. (1994). Acquiring and combining overlapping concepts. *Machine Learning*, *16*, 121–155.

McClosky, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24). San Diego, CA: Academic Press.

McKusick, K. B., & Langley, P. (1991). Constraints on tree structure in concept formation. *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 810–816). Sydney: Morgan Kaufmann.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, *18*, 203–226.

Müller, G. E., & Pilzecker, A. (1900). Experimentalle Beiträge zur Lehre vom Gedächtnis. *Zeitschrift fur Psychologie*, *1*, 1–300.

Neri, F., & Saitta, L. (1993). Exploiting example selection and ordering to speed up learning. *Proceedings of the AAAI Spring Symposium on Training Issues in Incremental Learning* (pp. 54–69). Stanford, CA: AAAI Press.

Porat, S., & Feldman, J. A. (1991). Learning automata from ordered examples. *Machine Learning*, *7*, 109–138.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Reigeluth, C. M.,, & Stein, F. S. (1983). The elaboration theory of instruction. In C. M. Reigeluth (Ed.), *Instructional design theories and models*. Hillsdale, NJ: Lawrence Erlbaum.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *2*, 285–308.

Rendell, L. A. (1986). A new basis for state-space learning systems and a successful implementation. *Artificial Intelligence*, *20*, 369–392.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: A production system model of practice. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three modes of learning. In J. W. Cotton & R. Klatzky (Eds.), *Semantic factors in cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Sammut, C., & Banerji, R. B. (1986). Learning concepts by asking questions. In R. S. Michalski,

J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2). San Mateo, CA: Morgan Kaufmann.

Schlimmer, J. C., & Fisher, D. (1986). A case study of incremental concept induction. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 496–501). Philadelphia, PA: Morgan Kaufmann.

Scott, P. D., & Markovitch, S. (1991). Representation generation in an exploratory learning system. In D. H. Fisher, M. J. Pazzani, & P. Langley (Eds.), *Concept formation: Knowledge and experience in unsupervised learning*. San Mateo, CA: Morgan Kaufmann.

Shapiro, A. D. (1987). *Structured induction for expert systems*. Wokingham, UK: Addison-Wesley.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

VanLehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence*, *31*, 1–40.

Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision.* New York: McGraw–Hill.