On the Necessity of U-Shaped Learning

Lorenzo Carlucci[1] and John Case[2]

[1] Department of Computer Science
Rome I University
Rome, Italy
carlucci@di.uniroma1.it
[2] Department of Computer and Information Sciences
University of Delaware
Newark, DE
USA
case@udel.edu

Abstract. A *U-shaped curve* in a cognitive-developmental trajectory refers to a three-step process: good performance followed by bad performance followed by good performance once again. U-shaped curves have been observed in a wide variety of cognitive-developmental and learning contexts. U-shaped learning seems to contradict the idea that learning is a monotonic, cumulative process and thus constitutes a challenge for competing theories of cognitive development and learning. U-shaped behaviour in language learning (in particular in learning English past tense) has become a central topic in the Cognitive Science debate about learning models. Antagonist models (e.g., connectionism vs. nativism) are often judged on their ability of modeling or accounting for U-shaped behaviour. The prior literature is mostly occupied with explaining *how* U-shaped behaviour occurs. Instead, we are interested in the *necessity* of this kind of apparently inefficient strategy. We present and discuss a body of results in the abstract mathematical setting of (extensions of) Gold-style computational learning theory addressing a mathematically precise version of the following question: Are there learning tasks that *require* U-shaped behaviour? All notions considered are *learning in the limit from positive data*. We present results about the necessity of U-shaped learning in classical models of learning as well as in models with bounds on the memory of the learner. The pattern emerges that, for parameterized, cognitively relevant learning criteria, beyond very few initial parameter values, U-shapes are *necessary* for full learning power! We discuss the possible relevance of the above results for the Cognitive Science debate about learning models as well as directions for future research.

1 Introduction and Motivation

A *U-shaped curve* in a cognitive-developmental trajectory refers to a three-step process: good performance followed by bad performance followed by good performance once again. In learning contexts, *U-shaped learning* is a behaviour in which the learner first learns the correct behaviour, then abandons the correct behaviour and finally returns to the correct behaviour once again. This kind of cognitive-developmental trajectory has been observed by cognitive and developmental psychologists in a variety of child development phenomena: language learning [9,60,90] understanding of temperature [90,91], understanding of weight conservation [9,90], object permanence [9,90], and face recognition [12].[3]

U-shaped curves in cognitive development seem to contradict the 'continuity model of cognitive development,' i.e., the idea that performance improves with age, and that learning is a monotonic, cumulative process of improvement.[4] Thus, the apparent regressions witnessed by U-shaped learning trajectories have become a challenge for competing theories of cognitive development in general and of language acquisition in particular.

[3] Siegler's [89] tracks the historical U-shaped *interest* in U-shaped behaviour, and, as a corollary, references some more developmental phenomena where U-shapes have been found.

[4] This concept of monotonic should not be confused with the interesting technical senses in computational learning theory [51,99]. These latter seem not to be cognitively relevant and, hence, are not dealt with herein.

The case of language acquisition is paradigmatic. In the case of the past tense of English verbs, it has been observed that, early in language acquisition, children learn correct syntactic forms (call/called, go/went), then undergo a period of ostensible over-regularization in which they attach regular verb endings such as 'ed' to the present tense forms even in the case of irregular verbs (break/breaked, speak/speaked), and finally reach a final phase in which they correctly handle both regular and irregular verbs.

This example of U-shaped learning has been used as evidence against domain-general associative learning theories of language acquisition by supporters of linguistic nativism. It has figured so prominently in the so-called 'Past Tense Debate' between connectionism and rule-based theory (the original papers are [87,76,79], but see [77,62] for a more recent follow-up) that U-shaped learning has become the test-bed for theories of language acquisition: competing models are often judged on their capacity of accounting for the phenomenon of U-shaped learning (see, e.g., [60,79,92]).[5]

The prior literature is typically concerned with modeling *how* humans achieve U-shaped behaviour. Instead, we are mostly interested in *why* humans exhibit this seemingly inefficient behaviour. Is it a mere harmless evolutionary accident or is it *necessary* for full human learning power — for humans being competitive in their genetic marketplace? This is of course presently empirically very difficult to answer. Herein we pursue, nonetheless for potential interesting insight into this problem, a *mathematically precise* version of the following question: are there some *formal* learning tasks for which U-shaped behaviour is *mathematically necessary*? We discuss a large and growing body of results ([15,13,5,14,28,23,24]) in the context of (extensions of) Gold's formal model of language learning from positive data [40] that suggest an answer to this latter question.[6] Gold's model has been very influential in theories of language acquisition [73,95,94,70,8] and has been developed and extended into an independent area of mathematical research [71,50].

The basics of the model are as follows. A learner is an algorithm for a (partial) computable function [83] that is fed an infinite sequence consisting of all and only the elements of a formal target language, in arbitrary order and possibly with repetitions. At each stage of the learning process, the learner outputs a corresponding hypothesis based of the evidence available so far. These hypotheses are candidate (formal) grammars for the target language. Learning in this context means that after some point the grammars produced by the learner are correct for the target language. Importantly, different criteria can be defined by imposing conditions on the cardinality of the set of correct grammars that the learner produces in the limit, combined with restrictions on the learner's memory and other computational power. In this context a U-shape occurs whenever, in the process of eventually successfully learning some language, a learner *semantically* returns to a previously abandoned correct conjecture.[7] For each learning criterion we consider, we say that a learner is a *non-U-shaped* learner if it commits no U-shapes while learning languages that it learns according to that criterion (i.e., as in the empirical settings, we mostly do not care about possible U-shapes on other languages[8]). We consider *non-U-shaped* learners since, mathematically, it is useful to examine the consequences for learning power when U-shapes are *forbidden*. U-shaped learning is *necessary* for a given criterion if some class of languages can be learned by that criterion, but not if one uses the same criterion — except with U-shapes forbidden.

[5] Interestingly, the idea that U-shaped cognitive development might be "the quintessential hallmark of the developmental process" has been advanced by Marcowitch and Lewcowicz in their short paper [58].

[6] Various extensions of Gold's model will hereinafter be referred to as just Gold's model.

[7] Semantic return can be to a syntactically very different grammar — but a grammar, nonetheless, for exact same language.

[8] The case of forbidding returning semantically to an abandoned conjecture on *all* languages is called *decisiveness* [71,38,5], and there is a little more about it below.

We present, then, results about the impact of forbidding U-shaped behaviour in a number of learning models/criteria within Gold's framework. In some cases of interest U-shaped learning will turn out to be unavoidable: if U-shapes are forbidden, strictly less classes of languages are learnable. The general pattern that so far emerges from this line of research is the following. For cognitively relevant, parameterized learning criteria, beyond very few initial parameter values, U-shapes *are necessary* for full learning power![9]

The paper is organized as follows. In Section 2 we review the main notions and assumptions of Gold-style learning theory and present our formal definition of U-shaped behaviour. In Section 3 we present results about the necessity of U-shaped learning in the context of classical learning criteria with no memory limitations. In Section 4 we present results about the necessity of U-shaped learning in the context of learning criteria with memory limitations. In Section 5 we discuss some features of the proof techniques that might be relevant for Cognitive Science. In Section 6 we present results about the necessity of forms of non-monotonic learning other than U-shaped learning. In Section 7 we offer a final discussion and prospects and seemingly difficult, cognitively relevant, open questions for future research.

2 Gold-Style Computational Learning Theory

We review the basic assumptions and ingredients of (extended) Gold-style computational learning theory [50] in an informal yet rigourous way.

2.1 Languages, Grammars, Texts and Learners

An *alphabet* is a finite set of symbols. A *language* is modeled as a set of finite strings from an alphabet. Without loss of generality — using standard coding techniques — a language can be identified with a set of natural numbers. This model of language may seem to be very naive but is broad enough to model most of the schemes of language description commonly used in Linguistics. Typically, the alphabet symbols are taken to represent the words of the language and the strings of alphabet symbols are taken to represent the sentences of the language. Other interpretations are equally possible: the elements of the alphabet can represent morphemes, phonemes, IPA symbols, and the elements of the language can be identified with the strings satisfying the phonotactical constraints of the language.

As in most computational learning theories, a learner in Gold's model is an agent that tries to identify a target language based on implicit information.

The learning process is modeled as an inductive procedure (indexed by a discrete time parameter n) in which a learner is trying to identify a target language based on implicit information. At time n the learner has to make a guess about what the target language is based on the finite amount of information seen so far. Gold's model is in this sense a theory of *inductive inference*.

In Gold's original model, and for all learning criteria of interest in the present paper, the information available to the learner is an infinite sequence consisting of all and only the elements of the language. The elements of the language can appear in any order whatsoever and with repetitions. Any such a sequence is called a *text* or a *presentation* of a language. Each (non-empty) language has infinitely many different presentations (indeed any infinite language has uncountably [42] many presentations). A learner is fed a text element by element, and after receiving each new piece of information, has to make a guess about the target language.

Children appear to be learning natural languages by a casual and unsystematic exposure to the linguistic activity of adults around them. The issue of whether children profit from negative

[9] We'll develop below some important, *parameterized* learning criteria.

information in language learning is still debated in Psycholinguistic (see, e.g., [59] and, for a review, the recent [36]).[10] A substantial body of experimental evidence suggests that children learn natural language in the *absence of feedback* ([11,59,92]). Accordingly, the learners in Gold's (to be sure idealized) model learn from *positive data* only: the learner receives as input all and only the correct sentences of the language. In [67] a well-documented body of experimental evidence suggests that children are *insensitive* to the way language is presented, in terms of order, repetitions, frequency and the like.[11] This is mirrored in Gold's model by the requirement that the learners correctly learn the language no matter how the input is presented, as long as it contains all and only the correct sentences of the language.

The hypotheses of a learning machine in Gold's model are (numerical codes for) computer programs in a pre-given (typically general) programming system. The idea here is that human language acquisition involves the acquisition of a *grammar* for the target language. How the knowledge of such a grammar is coded in the brain is still unknown. According to formal language theory, a *grammar* for a language is a finite list of rules that effectively *generate* all and only the correct sentences of the language. Such a grammar can be identified with a computer program such that, when the program is run, all and only the elements of the language are produced as output. Such a computer program has to be written in some programming language. Nowadays we have a mathematically precise notion of *algorithm* and of general-purpose (universal) programming language. We can thus fix a programming language in which any algorithm can be implemented and ask that the grammars are written in that language. Any general-purpose high level modern programming language will do. Formally, an *acceptable programming system* is a universal programming system (such as Turing Machines, Random Access Machines, C, Lisp, etc.) into which one can compile from any programming system, or, equivalently, in which any control structure can be implemented [82,83,85]. Computer programs are finite objects and can thus be coded as natural numbers (see [83]). We refer to these codes as *names* of the associated computer program and as *indices* of the corresponding language generated by the computer program.

What kinds of languages can be captured by such grammars? The Chomsky Hierarchy classifies formal languages in terms of the complexity of the grammars that generate them. The most general class is the class of computably enumerable (c.e.) languages. These are the languages that can be generated by arbitrary algorithmic procedures. Any such language has at least one index in any fixed acceptable programming system (indeed, it has infinitely many distinct ones, corresponding, for example, to the different ways of generating the language). The view that the class of natural languages could be identified with one of the classes of the Chomsky Hierarchy had been dominant in Cognitive Science for many years. Nowadays researchers are more inclined to think that natural languages form a class that is orthogonal to the classes of the Chomsky Hierarchy. It is clear that context-free languages are not enough to model all natural languages ([80,54,88,10,46]), but no one objects to the idea that natural languages are computably enumerable.

The modeling of all human cognition by algorithms or, equivalently, by computer programs is a well-established trend in Cognitive Science ([53,81]). Accordingly, we only consider those computing (partial) computable functions, i.e., learners whose behaviour can be simulated by an algorithm. On the other hand, each (non-empty) language admits many non-computable presentations.[12] It might be interesting to study U-shaped learning with the restriction to computable texts, but we do not make such an assumption here. Note that, for some learning

[10] Some theoretical results appear, e.g., in [4]. See below for more such references.

[11] Some associated theoretical results for some learning criteria are presented in [18].

[12] It can be argued that non-computable texts can be and are generated by randomness in the environment, including from quantum mechanical phenomena.

criteria, it is known that restricting to computable texts makes no difference as to successful learning [18].

2.2 Successful Learning

What does it mean that a learner *learns* a language? Empirically, we say that a child has acquired knowledge of a natural language after the time he or she stops to make errors (or if the error rate drops below a threshold) and starts to generalize (i.e., produce original linguistic output) correctly. We can (at least currently) never be sure that some error will not occur later on in an individual's linguistic behaviour. From a theoretical viewpoint, however, it makes sense to require that knowledge is acquired once a point is reached in the process of hypothesis formation, after which the learner does not make wrong guesses about the target language.

We do not ask that the learner *knows* when this point of convergence has been reached and consequently stops learning. The process of learning is *in the limit* in the sense that the successful learner will eventually output only correct conjectures but will not necessarily be able to know that he/she is doing so and consequently halt the learning process. To require this would result in serious limitation of the power of the model. We would here like to quote Gold's [40] own justification for studying learning in the limit.

> A person does not know when he is speaking a language correctly; there is always the possibility that he will find that his grammar contains an error. But we can guarantee that a child will eventually learn a natural language, even if it will not know when it is correct.

We further illustrate the setting and introduce some terminology. Suppose that, at some point, after reading the elements $t_0, t_1, \dots t_n$ of a text for a language L, machine (algorithmic procedure) $\mathbf{M}$ conjectures a grammar g_n, and that grammar g_n is a correct grammar for L. Suppose now that, after conjecturing g_n, $\mathbf{M}$ outputs *forever* only correct grammars for L, i.e., all later grammars $g_{n+1}, g_{n+2}, \dots$ output while seeing the rest of the input text $t_{n+1}, t_{n+2}, \dots$, are correct grammars for L. In that case we say that $\mathbf{M}$ has *converged* to a set of correct grammars, in this case to the set $\{g_n, g_{n+1}, g_{n+2}, \dots\}$. We call this set the set of $\mathbf{M}$'s *final conjectures*.

For all criteria of interest to the present paper, a learner is required to converge to *a set of correct grammars* for the target language, in response to *any* text for the target language. Different learning criteria can be defined by imposing conditions on the set of correct grammars to which the learner converges (e.g., the set is a singleton, the set has cardinality less or equal to than b, the set is finite, ...). Other criteria of interest are obtained by imposing restrictions on the learner's memory (plausible for us humans), and will be discussed later.

It is commonly assumed that children are able to learn any natural language, given the appropriate input. Any proposed definition of natural language determines *a class* of languages that fulfill the definition. Accordingly, we are interested in machines that learn *classes* of languages, rather than individual languages. In fact in Gold's model learning a single language is trivial: the learner can blindly output a fixed grammar for the language, regardless of its received input. A learner is said to learn a *class* of languages according to a given learning criterion if the learner learns every language in the class according to that criterion.

2.3 U-Shaped Behaviour

As Strauss and Stavy write in the Introduction to [90], U-shaped learning consists in "the appearance of a behaviour, a later dropping of it, and what appears to be its subsequent

reappearance [...] Phase 1 behaviour is a correct performance and Phase 2 is an incorrect performance, whereas Phase 3 behaviour is a correct performance." From a theoretical viewpoint we decide to interpret a *good performance* as the (sometimes observable) consequence of the learner's conjecturing a *correct grammar*, and a *bad performance* as the (sometimes observable) consequence of the learner's conjecturing a *wrong grammar*, i.e., a grammar for a language differing from the target language. This is a reasonable working hypothesis, and we adopt this perspective in our formal setting. It is indeed analogous to assuming that a child learning a language eventually acquires at least one grammar for the learned language.

We say that a learner learning a class $\mathcal{L}$ of languages is *U-shaped on the class* $\mathcal{L}$ if it exhibits U-shaped behaviour on *some* presentation of *some* language in the class. That is, a machine $\mathbf{M}$ is U-shaped on a class $\mathcal{L}$ of languages if there is a language L in $\mathcal{L}$ and a text T for L such that, while learning L from T, $\mathbf{M}$ outputs at some point a correct grammar for L, then later abandons it and makes a wrong conjecture, and later outputs a correct conjecture again. More formally, if the text T for L is the infinite sequence $t_0, t_1, t_2, \ldots$, $\mathbf{M}$ is *U-shaped on* T if there exists three elements t_m, t_n, t_p such that (1) $m < n < p$, and (2) after reading the input up through t_m the machine $\mathbf{M}$ conjectures a grammar g_m which is a correct for the language L, (3) after reading the input up through t_n the machine $\mathbf{M}$ conjectures a grammar g_n which is not a correct grammar for L, and (4) after reading the input up through t_p the machine $\mathbf{M}$ conjectures a grammar g_p which is again a correct grammar for L. We do *not* require the two correct conjectures (g_m and g_p) to be *syntactically* the same. We only ask that they both generate the target language L. We occasionally refer to this form of U-shaped learning as *semantic U-shaped learning*. For mathematical convenience we will state our results in terms of *non-U-shapedness*. A learner is said to be *non-U-shaped* on a class $\mathcal{L}$ of languages if and only if it is not U-shaped on the class. When such a learner $\mathbf{M}$ is presented with the elements of an L it learns in the order of some text T, if at some point $\mathbf{M}$ outputs a correct conjecture for L, then all conjectures output by $\mathbf{M}$ after that point are correct conjectures for L. As just hinted, we actually mostly care about U-shaped behaviour of learners on classes of languages they actually learn (according to some fixed learning criterion of interest), so that convergence on a set of correct conjectures is ensured from the onset for every text for every language in the class under consideration.

Our definition of U-shaped learning obviously constitutes an *idealization* of the experimentally observed learning behaviours. 'U-shaped learning' is an *experimental* concept. A U-shaped *learning curve* is a qualitative feature of a quantitative representation of measurable *linguistic performance* rather than of *linguistic competence*. At the present time it is impossible to look into people's heads to see what grammar(s) they are using. The same *current* empirical untestability holds for the problem of testing whether the grammar used in Phase 1 is the same as the grammar used in Phase 3. We thus believe that requiring correct identification of the target language in Phase 1 of a U-shaped curve is a viable abstraction — in fact akin to the common idea that a competent learner knows some grammar for the language. Also, we believe that biological biases and genetic constraints make the possibility of the child reaching a correct linguistic knowledge (i.e., a correct grammar) at an early age not so far-fetched.

We will occasionally mention stronger forms of non-U-shaped learning appearing in the literature. *Strong non-U-shaped learning* (from [97], where it is called *semantically finite*) refers to learning in which the stronger requirement of never *syntactically* returning to a previously abandoned conjecture is imposed. We also call it *syntactical* non-U-shaped learning. Another strong notion appears in the literature: as above, a *decisive* learner [71] is a learner that never *semantically* returns to a previously abandoned conjecture — be it right or wrong.

We are now able to formulate, for each formal learning criterion, the following fundamental questions. Is U-shaped behaviour necessary for the full learning power? Are there classes that

are learnable *only* by resorting to U-shaped behaviour (on some text for some language of the class)?

3 U-Shaped Learning with Full Memory

In this Section we present results about the necessity of U-shapes in three classical learning contexts in which the learner has full access to previously seen data items. At any stage n of the learning process, such a learner can access the full initial segment $t_0, t_1, \ldots, t_n$ seen up to n of the input text T and consequently recompute any of its own previously output conjectures $g_0, g_1, \ldots, g_{n-1}$ as well as the new one g_n. *Explanatory Learning* requires the learner to converge to *a single correct hypothesis in the limit*. Explanatory Learning is Gold's [40] original model of learning in the limit. At the other extreme, *Behaviourally Correct learning* [26,72] allows the learner to stabilize on *possibly infinitely many syntactically different correct conjectures in the limit*. Vacillatory Learning is intermediate between the two: the learner there is allowed to vacillate between *at most a finite fixed (or finite unbounded) number of correct conjectures in the limit*.[13] Vacillatory Learning [18] defines an infinite hierarchy of more and more powerful learning criteria strictly intermediate between Explanatory and Behaviourally Correct Learning. The case of Vacillatory Learning is paradigmatic and is the first example of a general pattern of which we will see more: in parametrized learning models, beyond very few initial parameters, U-shaped learning is necessary for full learning power.

3.1 Explanatory, Behaviourally Correct and Vacillatory Learning

We here more formally define Behaviourally Correct, Explanatory and Vacillatory Learning.

The minimal requirement for a learner $\mathbf{M}$ to learn a language L is that $\mathbf{M}$, given any text for L, eventually outputs only correct grammars for L. These grammars can possibly be infinitely many syntactically distinct ones. A learner that satisfies this minimal and correct convergence requirement is said to *behaviourally* identify the language L. Such a learner identifies the language only *extensionally* but not *intensionally*: it does not have to stabilize on a single grammar for the language, but is nevertheless able to correctly capture the *extension* of the language and thus to eventually reach correct linguistic *behaviour*. We refer to this criterion as *Behaviourally Correct Learning*.

By fixing a finite number $b \geq 1$ as an upper bound to the number of correct conjectures to which a learner is allowed to converge in the limit we obtain the concept of *Vacillatory Learning* with vacillation bound b. For each choice of a positive natural number b we get a distinct criterion. If we require that the number of correct conjectures to which the learner converges is finite (but undetermined), we get a different criterion.

Finally, if we require that the learner converges to *a single correct grammar*, we have the concept of *Explanatory Learning*. This is Gold's original concept from [40], and is obviously equivalent to vacillatory learning with vacillation bound of $b = 1$. Such a learner is said to *explanatory* learn the language in the sense of stabilizing on a single correct *description* or *explanatory definition* of the language — without changing its mind later.

We compare the power of the learning criteria by comparing the classes of learnable languages as sets. A learning criterion is *more powerful* than another if it allows learning of more classes of languages. What is learnable in the explanatory sense is also learnable in the vacillatory and in the behaviourally correct sense — by definition. For each b, learning with vacillation bound

[13] An example fixed finite bound would be 3. Finite *un*bounded allows different finite bounds (on the number of successful programs in the limit) on different texts and languages (on which the learner is successful).

b is contained in learning with vacillation bound $b+1$ by definition and in learning with an arbitrary but finite number of correct grammars in the limit.

It is known that all the above mentioned inclusions are indeed *strict* [18]. Thus, all the learning criteria defined so far are different and give rise to a hierarchy of more and more powerful learning paradigms — according to which more and more language classes become learnable.

It might be profitable here to recall that the general problem of deciding whether two grammars generate the same language is algorithmically undecidable. Also, grammars for the same language can be so different that it is impossible to prove their equivalence from the axioms of Set Theory [83]. This gives a hint on what can be gained by allowing a learner more than one correct conjecture in the limit.

The Vacillatory Learning criteria with vacillation bounds $b = 1, 2, 3, \ldots$ form an infinite strict *hierarchy* of more and more powerful learning criteria, on top of which is learning with an arbitrary finite number of correct conjectures in the limit [18]. We call this hierarchy the *Vacillation Hierarchy*. We state the existence of this hierarchy as a theorem for ease of further reference.

Theorem 1. *For each choice of a positive natural number b there are classes of languages that are learnable by vacillating between at most $b+1$ correct grammars in the limit but not by vacillating between at most b grammars. Also, there are classes that are learnable by converging to a finite but not preassigned number of correct grammars but are not learnable by vacillation between at most b correct grammars, for any choice of a positive natural number b.*

Let us briefly describe in detail a class of languages that witnesses the separation between learning with vacillation bound $b+1$ and b.[14] Since grammars are coded as natural numbers and languages are sets of natural numbers, it might well be the case that a (code for a) grammar g generating a language L is also an element of L. Choose a positive natural number b. Now consider the class $\mathcal{L}_{b+1}$ of languages defined by the following two requirements: (1) Among the first $b+1$ elements of L (with respect to the usual order $<$ of the natural numbers) there occurs at least a code of a grammar for L, and (2) None of the elements of L beyond the $b+1$-th is a code of a grammar for L. Consider, for example, the concrete case of $b = 1$. In other words, $\mathcal{L}_2$ contains all and only those languages L such that one or both of the first two elements of L, and no other, is a code for a grammar for L. A variant of a proof of a theorem of [18] shows that the class $\mathcal{L}_{b+1}$ *can* be learned by vacillating in the limit between at most $b+1$ correct grammars, but *cannot* be learned by vacillating between at most b correct grammars. Thus the class $\mathcal{L}_{b+1}$ witnesses the fact that learning with vacillation bound $b+1$ is *strictly more powerful* than learning with vacillation bound b.[15] The results can be strengthened by showing that even allowing a learner to converge on approximately correct conjectures (in the sense of grammars that identify the target language up to a finite number of errors) does not make the class $\mathcal{L}_{b+1}$ learnable with vacillation bound b [18].

3.2 U-Shapes in Explanatory, Vacillatory and Behaviourally Correct Learning

The first result in the area [5] showed that U-shaped learning is *not* necessary for the full power of Explanatory Learning, as stated in the following Theorem.

Theorem 2. *Every class of languages that can be learned by convergence to a single correct grammar can be learned in this sense by a non-U-shaped learner.*

[14] This class is a variant of that employed in [18] and is from [14].

[15] A similar class shows that learning with arbitrary but finitely many correct grammars in the limit is strictly more powerful than learning with vacillation bound b for each positive natural number.

A very succinct proof can be found in the more recent [24]. The result has been strengthened in [27] to show that any explanatory learner can be transformed into an explanatory learner that is strongly non-U-shaped. This has to be contrasted with the fact that decisive learning — i.e., imposing no semantical return to *any* previously abandoned conjecture — *does* restrict learning power of explanatory learners — as shown in [5].

From a Cognitive Science perspective the above results means that, *if* Explanatory Learning as such is an adequate model of human learning acquisition, *then* U-shaped behaviour is an *un*necessary feature of human behaviour. However, few may be inclined to think that Explanatory Learning is an adequate model of human language acquisition. In particular, the requirement that the learner must converge on exactly one correct grammar in the limit seems to be too restrictive. While it is possible to measure changes in linguistic *behaviour* experimentally, as noted above, it is not currently possible to detect experimentally syntactic changes in people's heads. Recall that, on the other hand, general grammar equivalence is algorithmically undecidable. Humans might be taking advantage of not committing to a single description of the target language.

Behaviourally Correct Learning is known to be strictly more powerful than Explanatory Learning, and it is interesting to investigate what the impact of U-shaped learning is in this context. As observed in [5,14], based on a proof from [38], U-shaped behaviour *is* necessary for the full learning power of behavioural learners, as stated in the following Theorem (see [14]).

Theorem 3. *There are classes of languages that can be behaviourally identified but cannot be behaviourally identified by a non-U-shaped learner.*

If Explanatory Learning seems to be too restrictive, Behaviourally Correct Learning seems much too liberal — in allowing the learner to converge on up to *infinitely* many distinct correct grammars for the target language. The case of infinitely many distinct correct grammars must include humanly unrealistically large size grammars.

Vacillatory Learning is more realistic in this respect, and gives rise to a completely different and much richer picture.[16]

We already know from Theorem 2 that U-shaped behaviour is redundant for the first level of the Vacillation Hierarchy (since it's just Explanatory Learning). What about the other levels, the levels in which the power of vacillation is actually used?

Consider a learner $\mathbf{M}_1$ learning a class $\mathcal{C}$ with vacillation bound $b > 1$. Suppose now that this learner is non-U-shaped on that class. Now imagine another learner $\mathbf{M}_2$ observing the behaviour of $\mathbf{M}_1$ on a text T for some language L in the class $\mathcal{C}$ and acting as follows. As soon as $\mathbf{M}_1$ outputs a conjecture *for the first time*, $\mathbf{M}_2$ outputs the same grammar. Each time $\mathbf{M}_1$ (syntactically) repeats a previously output conjecture, $\mathbf{M}_2$ just forgets it, and outputs again its own most recent conjecture. We claim that $\mathbf{M}_2$ learns the language L in the explanatory sense, i.e., that $\mathbf{M}_2$ will eventually converge on a single correct conjecture for L. Why is that so? Since $\mathbf{M}_1$ learns L in the vacillatory sense, at some point n of the learning process, $\mathbf{M}_1$ will output, *for the first time*, a correct conjecture for L, call it g. By design, $\mathbf{M}_2$ will output g as well. Since $\mathbf{M}_1$ by assumption converges on at most b different correct grammars, $\mathbf{M}_1$ can output after g only *finitely many previously unseen* grammars. Also, since $\mathbf{M}_1$ is by assumption non-U-shaped on L, all conjectures output by $\mathbf{M}_1$ after g are *correct* conjectures for L.

Let g' be the — last but not least! — grammar to appear in $\mathbf{M}_1$'s output beyond g, i.e., *after all other grammars output by* $\mathbf{M}_1$ *from* g *on have already been output once.* It is easy to see

[16] For the *human case*, the b, bounding the number of correct grammars in the limit, must have an upper limit. A slight paraphrase of an relevant argument from [18] follows. At least one of b distinct grammars would have to be of size proportional to the size of b (i.e., to $\log b$); hence, for extraordinarily *large* b, at least one of b distinct grammars would be too large to fit in our heads — unless, as seems highly unlikely, human memory storage mechanisms admit infinite regress.

that $\mathbf{M}_2$ will converge on g'. Thus, $\mathbf{M}_2$ will learn L in the explanatory sense. As L and T were arbitrary in the above argument, it shows that $\mathbf{M}_2$ learns the whole class $\mathcal{C}$ in the explanatory sense.

So we have the following Theorem from [14].

Theorem 4. *For all $b > 1$ the following holds. Every class that is learnable by vacillation between at most b correct grammars and without U-shapes is already learnable by convergence to a single correct grammar.*

U-shaped behaviour is therefore necessary for the full power of Vacillatory Learning in a very strong sense: if U-shapes are forbidden, then the extra power gained by vacillation is lost. The Vacillation Hierarchy (see Theorem 1) collapses to Explanatory Learning.

It is an easy consequence of Theorems 2, 1, and 4 that the class $\mathcal{L}_{b+1}$ (defined above) cannot be learned without U-shaped behaviour by vacillating between at most $b+1$ grammars in the limit. In fact, the same holds if the learner is allowed to converge on grammars approximating the target language modulo a finite number of anomalies. Consider the particular case of $b = 1$. Then $\mathcal{L}_2$ is *a concrete example* of a class that *requires* U-shaped behaviour to be learned. In fact we know from above that $\mathcal{L}_2$ is learnable by vacillating between at most two correct grammars but not by converging to a single correct grammar in the limit. Consider a learner $\mathbf{M}$ learning $\mathcal{L}_2$ by vacillation between no more than 2 correct grammars. Such a learner exists because $\mathcal{L}_2$ is learnable with vacillation bound 2. Now suppose that $\mathbf{M}$ does *not* exhibit U-shaped behaviour. By the argument sketched above (see Theorem 4), this would mean that there is another learner, $\mathbf{M}'$ that learns the class $\mathcal{L}_2$ in the explanatory sense. But this is a contradiction. Formally, we have the following Corollary (see [14]).

Corollary 1. *Let $b \in \{2, 3, \ldots\}$. Then any $\mathbf{M}$ witnessing that $\mathcal{L}_b$ is learnable with vacillation bound b necessarily employs U-shaped learning behaviour on $\mathcal{L}_b$.*

The above results suggest that, *if* some of the natural language learning tasks humans have to face are of the kind of the classes $\mathcal{L}_{b+1}$, and *to the extent that* Vacillatory Learning is an adequate model of human language learning, *then* U-shaped behaviour is not an harmless and accidental feature of human behaviour, but may be *necessary* for learning what humans need to learn to be competitive in their genetic marketplace.[17] From this perspective it would be interesting to find insightful *characterizations* of the language classes that require U-shaped behaviour to be successfully learned in the vacillatory sense.

3.3 Getting Around U-shapes

We know from the previous section that some classes require U-shaped behaviour in order to be learned in the vacillatory sense. But how *deep* is the necessity of U-shaped behaviour in such cases? What happens if we remove the finite vacillation bound and consider behaviourally correct learnability of those classes? Obviously every class which is vacillatorily learnable is also behaviourally learnable by definition. But can we avoid some U-shapes if we only have to learn in a behaviourally correct way? The picture that emerges is interesting in our opinion. First, if we consider behaviourally correct learnability of the classes in the first non-trivial level of the Vacillation Hierarchy the necessity U-shaped behaviour disappears. We have the following Theorem from [14].

Theorem 5. *Every class that can be identified by vacillating between at most 2 indices can also be behaviourally identified by a non-U-shaped learner.*

[17] See the discussion in Section 5.1 below re the thesis from [18] that self-referential examples portend natural ones.

In contrast, from *the third* level of the Vacillation Hierarchy on the necessity of U-shaped learning cannot be removed even by allowing the learner to converge to infinitely many syntactically different correct grammars in the limit! We have the following Theorem from [14].

Theorem 6. *For $b > 2$ there are classes that can be identified by vacillating between at most b correct grammars but which cannot be behaviourally identified by any non-U-shaped learner.*

In this sense, we can say that the necessity of U-shaped behaviour for these classes is even *deeper* than the necessity of U-shaped behaviour for learning with vacillation bound 2. Note how a (arguably) cognitively implausible model of learning such as Behaviourally Correct Learning can be usefully used to qualitatively strengthen results about the un-eliminability of a cognitively relevant learning strategy (U-shaped learning) for some not so implausible learning model (Vacillatory Learning). It is difficult to judge whether the asymmetry between 2 and strictly larger than 2 might have some significance from the Cognitive Science perspective.

4 U-Shaped Learning with Memory Limitations

For modeling humans, a major limitation of the models considered so far is that they allow a learner too easy access to *all* the previous data. Humans certainly have memory limitations. It is therefore of cognitive science relevance to investigate the impact of forbidding U-shaped strategies *in the presence of memory limitations*. In this Section we present results about the necessity of U-shaped learning in learning models featuring *very severe* memory limitations. For all these models, the convergence requirement is the same as for Explanatory Learning: a single correct grammar must be output in the limit. The models differ in the forms of memory allowed to a learner.

It is profitable to distinguish between intensional memory and extensional memory, although they cannot always be kept distinct. *Intensional memory* refers to the learner's memory of his own past conjectures. *Extensional memory* refers to the learner's memory of previously seen data items.

4.1 Iterative Learning

Iterative Learning [96,95] is a fundamental model of inductive learning with memory limitations. An iterative learner computes its guesses about the target language based on its own *most recent* conjecture and on the *current* data item *only*.[18] Iterative Learning is a well-studied model [56,21,49,47,29,6]. Most interestingly, from the perspective of the present paper, Iterative Learning is the base case of a hierarchy of stronger and stronger memory-limited models. It is thus an interesting question whether U-shapes are necessary or not in this model. An attempt at answering this question was made in [13]. The problem was solved only later by Case and Moelius [28]. Their result shows that U-shapes are *not* necessary for the full learning power of iterative learners.

Theorem 7. *All classes of languages that can be learned by an iterative learner can be learned by an iterative learner without U-shapes.*

By contrast with Theorem 7 just above, it is shown in [24] that some iteratively learned classes can*not* be iteratively learned *strongly* non-U-shapedly.

[18] Note that an iterative learner can make up from its inability to explicitly remember previously seen data items by *coding* them into its conjectures. This trick can only be used a finite number of times, and has to stop by the time the learner has converged to its final conjecture. Knowing when to stop is, of course, a hard part.

4.2 Bounded Example Memory Learning

Few would defend the claim that humans are iterative learners. At the very least, humans have some form(s) of extensional memory. Yet Theorem 7 above is important for the investigation of U-shaped learning in memory-limited models. As observed above, Iterative Learning is the base case of a parametrized family of criteria of learning with bounded memory of past examples. It is indeed easy to extend Iterative Learning by allowing learners to store a bounded number of items in their long-term memory. A Bounded Example Memory learner with memory bounded by n is an iterative learner that can store in memory up to n previously seen data items. This model has been introduced in [57] and studied further in [21]. A hierarchy of more and more powerful learning criteria — the Bounded Example Memory Hierarchy — is obtained by increasing the size of the long-term memory: for every $n \geq 0$, with the storing of $n+1$ data items in memory, more language classes can be learned than by storing only n data items. If a learner is allowed to store an arbitrary but finite number of items in its long-term memory, a criterion is obtained that is strictly more powerful than each finite level of the hierarchy. Allowing long term memory of one previously seen data item is already strictly stronger than Iterative Learning. The impact of forbidding U-shaped behaviour in this setting is largely unknown and is of primary interest for future research! Mathematically, it seems interestingly very difficult.

We strongly conjecture that U-shapes are necessary in the Bounded Example Memory hierarchy — at least beyond the first few memory bound parameter values!

Below we present results on models with *more severe* memory limitations.

4.3 Bounded Memory States Learning

In *Bounded Memory States Learning* [13] a learner has an explicit bound on it's memory and otherwise only knows it's current datum. *No* access to previously seen data and to previously formulated conjectures is allowed![19] At each step the learner computes its conjecture as a function of the current (bounded) memory state and the current data item. The learner also chooses the new (bounded) memory state to pass on to the next learning step. Intuitively, for $c \geq 1$, a learner that can choose between c memory states is a learner that can store one out of c different values in its memory. When $c = 2^k$, a learner with c memory states is equivalent to a learner with k bits of memory. Bounded Memory State learning with an arbitrary but finite number of memory states is equivalent to Iterative Learning (see [13]).

It was shown in [13] that U-shaped behaviour does *not* enhance the learning power of bounded memory states learners with only two memory states. Note that two memory states amount to one bit of memory. The full picture was later obtained by Case and Kötzing in [23]. This gives another instance of the 2 vs. 3 phenomenon.

Theorem 8. *There are language classes that are learnable with three memory states but cannot be learned without U-shaped behaviour with any finite number of memory states.*

The above result is consistent with the emerging picture so far: U-shaped learning is unavoidable in parametrized learning models beyond a few initial parameters. On the other hand, U-shapes are *un*necessary for Bounded Memory States learning with an arbitrary but finite number of memory states. This was proved in [23] on the bases of Theorem 7 and of the fact that the model is equivalent to Iterative Learning. Again, the limit case (arbitrary but finitely many) behaves very differently from the finite cases. The same might be the case for Bounded Example Memory Learning.

[19] In the human case it is plausible that we *do* have available our prior working hypothesis for computing our next one. This makes it more urgent to solve the problem of whether U-shapes are necessary (beyond the first few parameter values) in the Bounded Example Memory Hierarchy!

4.4 Memoryless Learning with Queries

Queries are meant to formalize a kind of interactive memory. A query is a question of the form "Have I previously seen the following item(s)?"

In *Memoryless Feedback Learning* [13] a learner may ask, in each round, a bounded number of queries about whether computed items have been previously seen in its input data — and otherwise only knows it's current datum. In this model the queries are *parallel*, in the sense, in each round, that the choice of a question — within each learning step — can*not* depend on the answer to a previous question. If *sequential* queries are allowed (each computed query beyond the first one can depend on the answer to the previous queries) we obtain the model of *Memoryless Recall Learning*, introduced in [23].

U-shaped learning is necessary for the full learning power of n-memoryless feedback learners, for every $n > 0$.

Theorem 9. *For every $n > 0$, there are classes of languages that can be learned with n parallel queries by a memoryless feedback learner but not with $n + 1$ parallel queries by a non-U-shaped memoryless feedback learner.*

As an open problem, it was asked in [13] whether this necessity could be overturned by allowing more queries. Is it the case that for every $m > 0$ there exists an $n > m$ (possibly with n much larger than m) such that all classes learnable with m queries can be learned with n queries but, then, without U-shapes? The question was answered negatively by Case and Kötzing in [23]. Indeed, much more was shown.

Theorem 10. *There is a class learnable with a single feedback query by a memoryless learner that cannot be learned by a non-U-shaped memoryless learner with any finite number of feedback queries, even if sequential (rather than parallel) queries are allowed.*

Interestingly, the above result is complemented by a result showing that any class of *infinite* languages that is learnable by a memoryless feedback learner with finitely many feedback queries is so learnable *without U-shapes*. In fact, all classes of infinite languages learnable with complete memory and, moreover, explanatorily learnable, can be learned without U-shapes by a memoryless feedback learner using a finite but unbounded number of feedback queries. We will see more about this pattern: the necessity of U-shaped learning *sometimes* disappears when learning classes of infinite languages only.

On the other hand, there is a class of infinite languages that can be learned by a memoryless feedback learner with a single feedback query, but that cannot be learned without U-shapes using any *particular* number of feedback queries.

It's a cognitively important open question whether U-shapes are necessary for feedback or recall learning for which the learner also knows its just prior working hypothesis/conjecture! This too seems to be interestingly mathematically difficult. From [21] it is known this kind of learning forms a hierarchy in dependence on the bound on the number of queries.

Suppose $m + n > 0$. A more general, cognitively important open question is whether U-shapes are necessary for full learning power for learning criterion where the learner knows its prior conjecture and its current input datum, can remember m prior data items and can (feedback or recall) query in each round n computed items. We conjecture that, at least for $m + n$ beyond the first few positive values, U-shapes *are* necessary!

4.5 Counters and Time/Data Awareness

Memory-limited learners such as human beings might take advantage from other forms of information during their training. For example, humans are to some large extent aware of the passage

of time and data — certainly when they are awake. Can this awareness give some additional learning advantage — in bounded memory cases?

In [28] the authors introduced an extension of Iterative Learning featuring the use of *counters*. In particular, they considered a model in which an iterative learner also knows *the number of not necessarily distinct data items it's seen so far*. In other words, the learner is aware of the iteration stage number of the learning process. This information is naturally coded as a *counter* going from 0 to infinity. We call this type of counter a *full counter*. We think of counters as modeling the fact that humans are at least somewhat aware of time and/or data passage. Some people may be more aware than others.

It was shown in [28] that iterative learners with full counters are strictly more powerful than plain iterative learners. Not surprisingly, they are also strictly less powerful than explanatory learners — since explanatory learners have available the data input sequence up to any point in time, so they can calculate the length of this sequence.

Kötzing [55] began a systematic study of *how* counters can improve learning power. Which properties of counters give a learning advantage? Is it the higher and higher counter values, which can, for example, be used to time-bound computations? Is it merely unbounded counter values? Is it, as above, knowing the exact number of data (not necessarily distinct) items seen so far? In [55] the impact of six different types of counters — each one modeling one of six potential advantages of using a counter — is fully studied — at least in the context of Iterative Learning. It is not so clear to us *yet* which type of counter best models human performance — again there may be human individual differences. More work needs to be done on this.

On the one hand, [55] showed that even the weakest of his six types of counter does improve learning power — at least of iterative learners. On the other hand, the six types of counters studied in [55] turned out to fall into two groups with respect to *iterative* learning power. The strongest learning advantage is given by having a full counter, but a strictly monotone one is also sufficient. Indeed, the proofs (again for Iterative Learning) show that the such a learner only needs to count the number of (not necessarily distinct) elements seen since it's last mind-change (i.e., change of conjecture) — or overestimate that number (however badly) to attain maximal learning power of using full counters. Dropping the monotonicity requirement already results in a loss of learning power (for Iterative Learning). The same learning power of an iterative learner using a not necessarily strict monotone but unbounded counter is achieved by dropping the monotonicity requirement *or* by using a counter that eventually enumerates all natural numbers but with no constraint on the order of the enumeration. More work needs to done on the learning advantages of various humanly plausible types of counters for learning criteria more humanly plausible than the very restrictive Iterative Learning.

Even the question of whether U-shaped learning is necessary for iterative learners with counters is an interesting still open problem for future research. In [28] the authors conjectured that U-shaped learning should be *unnecessary* if the learner has access to a full counter.[20] The conjecture is still open. [55] contains preliminary results on the interplay between counters and U-shaped learning in the context of a toy model of learning vastly weaker than even Iterative Learning: *Transductive Learning* is learning with no memory at all. A transductive learner outputs its new conjecture based on the current datum only. It is shown in [55] that in the context of Transductive Learning its six counter types give rise to four distinct extensions of Transductive Learning. For Transductive Learning, U-shaped learning exhibits a sensitivity to the counter being ordered: it makes a difference whether the learner has access to a monotone

[20] Interestingly, the conjecture is based on the observation that in the case of Iterative Learning, the dispensability of U-shapes in learning classes consisting of infinite language only could be much more easily proved than in the case of mixed classes. By analogy, *perhaps* the access to some form of infinity — as given e.g., by an unbounded counter — could make U-shapes unnecessary.

and unbounded counter rather than just to an unbounded counter (while it made no difference for Iterative Learning).

Here is a collection of master open questions (so far) regarding the necessity of U-shaped learning for humanly plausible criteria with memory limitations.

Suppose $m+n>0$. Are U-shapes necessary for full learning power for the learning criterion where the learner knows its prior conjecture and its current input datum, can remember m prior data items and can (feedback or recall) query in each round n computed items *and* has access to one of several humanly plausible counters? We conjecture that, at least for $m+n$ beyond the first few values and for some reasonable such counters, U-shapes *are* necessary!

5 On the Proof Techniques

Do *the proofs* of the above results give us any insights into the human case of necessity/indispensability of U-shaped behaviour? In this Section we discuss some features of some the proofs of results from previous sections that might be of interest from the Cognitive Science perspective.

5.1 Self-Reference, Self-Description and Self-Learning

Self-reference is a powerful technique in Computability Theory and figures prominently in Computability-theoretic Learning Theory as a whole. Many theorems proving the necessity of U-shaped learning make use of self-referential algorithms and language classes. The classes of languages witnessing many of the separations between a learning criterion and its non-U-shaped variant are *self-describing classes*. A self-describing class of languages is such that information about the grammars for the languages in the class is directly present in some elements of the languages themselves. For each $b \geq 2$, the classes $\mathcal{L}_b$ used in Section 3 to separate Vacillatory Learning with bound b from its non-U-shaped variant are self-describing classes: each language in $\mathcal{L}_b$ contains information about its own (coded) grammar within its first $b+1$ elements.

There has been a trend from [30,26,31,17] to [18] to [21] for the self-describing classes employed to go from obvious choices to very difficult choices, yet the original reason for their employment was to make the positive half of proofs (that one criterion has more power than another) immediate.

Recently, the use of self-describing classes has been generalized, in many cases improved (as regards immediacy of such positive halves of proofs), and systematized by Case and Kötzing [23,24,25,55]. For a given learning criterion $\mathbf{C}$, the *self-learning class* of languages *for* $\mathbf{C}$ is that class that is $\mathbf{C}$-learned by merely treating the data elements as (codes for) programs to be run on inputs relevant to $\mathbf{C}$-learning. This may seem irrational, and, of course, many numbers so run as programs will produce no output (conjectures). Such numbers will, then, not be data elements of languages in the self-learning class for $\mathbf{C}$! Surprisingly it has been shown very generally (including for criteria as above) that the self-learning class for $\mathbf{C}$ witnesses that $\mathbf{C}$ is more powerful than another criterion $\mathbf{C}'$ if and only if this is indeed the case! In particular, then, the self-learning class for $\mathbf{C}$ will witness $\mathbf{C}$-learning is more powerful than the variant of $\mathbf{C}$ in which U-shapes are forbidden if and only if U-shapes are *necessary* for full $\mathbf{C}$-learning power!

The technical counterparts of the use of self-describing and self-learnable classes are the so-called Recursion Theorems of Computability Theory. The first such theorem is due to Kleene (see e.g., [83, page 214]) and can be stated as follows. Let p be an arbitrary algorithmic task. Then there exists an program e (depending on p) that acts as follows. When run on an input x, the program e creates a copy of its own code (a self-copy) and performs the task p on the combined input consisting of the self-copy and the external input x. That is, e performs the preassigned algorithmic task p on its own code combined with the external input x. In some

important sense e exhibits a form of *usable* self-knowledge (in this case the ability to produce and manipulate a low-level self-description). Actually, the self-knowing program e can be found algorithmically from p — and quite efficiently so [86]. A number of extensions of Kleene's Recursion Theorem are known. One of the most far-reaching is Case's Operator Recursion Theorem [16]. This theorem features *infinitary self-* and *other-* reference, and it is important for the employment of self-learning classes.

Even though (imperfect) self-knowledge and self-description might be powerful resources in the case of human learning, some may be dissatisfied by the fact that the classes of languages separating a learning criterion from its non-U-shaped variant are self-referential classes (or self-describing, or self-learning), rather than more *natural* classes of languages.

An analogy can be drawn between the self-referential witnesses of separations in Learning Theory and the original self-referential witnesses of Gödel's Incompleteness phenomenon [39] in formal systems of mathematics. The relevance of Gödel's First Incompleteness Theorem for classical mathematics was questioned on the basis of the fact that *his* unprovable and irrefutable sentences witnessing the incompleteness of formal systems (such as Peano Arithmetic and Zermelo-Fraenkel Set Theory) were self-referential sentences — sentences whose mathematical content was devoid of interest for non-logician mathematicians. It took more than forty years to find 'natural' examples of Gödel sentences. The first such example is the famous 'Large Ramsey Theorem' of Paris and Harrington [43]. Later, perfectly natural mathematical theorems such as Kruskal's Tree Theorem and the famous Robertson's and Seymour's Graph Minor Theorem were proved to require unexpectedly strong axioms. Partly based on this analogy, Case [18] proposed the following Informal Thesis.[21]

> Informal Thesis [18]: If a self-referential example witnesses the existence of a phenomenon, there are *natural* examples witnessing the same!

The other basis for the just above Informal Thesis is that self-reference arguments lay bare *an* underlying *simplest* reason for the theorems they prove [83,17,18,25]; if a theorem is true for such a simple reason, the "space" of reasons for its truth may be broad enough to admit natural examples.

We consider it nonetheless a difficult and interesting challenge for future research to find humanly natural examples of classes *requiring* U-shaped learning. This is in part because much of human cognition is hidden in brains too complicated for current experimental techniques.

5.2 The Problem of Generalization: Rules and Exceptions

According to a classical explanation of U-shaped learning, U-shapes occur because of the learner adopting two learning strategies: memorization of a finite table and production of a general rule [74,75,60]. This explanation has been most notably challenged by connectionists, who posit a single learning mechanism.

It is interesting to note how advocates of opponent theories agree in describing the learning situations in which U-shaped learning occurs as critically featuring an interplay between a small set of *exceptions* and a *general rule* or common case.

[21] Johnson [52] makes a similar analogy between Gold's Theorem (providing the unlearnability of the class of regular languages) and Gödel's Theorem: "In general, the relation of Gold's Theorem to normal child language acquisition is analogous to the relation between Gödel's first incompleteness theorem and the production of calculators. Gödels' theorem show that no accurate calculator can compute every arithmetic truth. But actual calculators don't experience difficulties from this fact, since the unprovable statements are far enough away from normal operations that they don't appear in real life situations." From this analogy Johnson concludes that Gold's Theorem is irrelevant to cognitive science, just as Gödel's Theorem is to concrete mathematics, apparently disregarding forty years of mathematical research on natural incompletenesses (see [37]).

Melissa Bowerman [9] thus points out that U-shaped learning curves "[...] occur in situations where there is a general rule that applies to most cases, but in which there are also a limited number of irregular instances that violate the rule," and goes on concluding that "[...] the solution involves a general adherence to the rule plus memorization of the exceptions." In the paper [84] of the connectionist school, the emergence of U-shaped behaviour is linked to learning tasks consisting of regularities (statistical ones in this case) and exceptions: "Specifically, we suggest that U-shaped curves can arise within a domain-general learning mechanism as it slowly masters a domain characterized by statistical regularities and exceptions."

An interplay between finite tables and infinite languages subsuming them *is intriguingly featured in some of our proofs*, e.g., by the proofs of Theorems 3 and 6 (see [14] for details).

The idea of these proofs is the following. Take a machine $\mathbf{M}$, and consider the behaviour of $\mathbf{M}$ on finite amounts of data. Consider the case of a finite sequence of data σ such that, when $\mathbf{M}$ receives σ as input, $\mathbf{M}$ outputs a grammar g such that the language generated by g is a proper superset of the elements of σ (i.e., it contains all the elements of σ *and some more*). Now suppose that $\mathbf{M}$ is required to learn *both* the finite language consisting of the elements of the sequence σ (call this language L_σ) *and* the language L_g generated by g. Consider now the behaviour of $\mathbf{M}$ on the following text. $\mathbf{M}$ is first fed σ. At this point, by choice of σ, $\mathbf{M}$ outputs grammar g for L_g. After that, $\mathbf{M}$ is presented with the elements of σ in any order. Since $\mathbf{M}$ must learn L_σ, at some point $\mathbf{M}$ will output a grammar generating the language L_σ, which is different from L_g. But since L_g contains L_σ, we can go on by presenting $\mathbf{M}$ with elements from L_g. Since $\mathbf{M}$ learns L_g by assumption, at some point $\mathbf{M}$ will output a correct grammar for L_g. But then $\mathbf{M}$ has committed a U-shape in learning L_g: it has first output a conjecture for L_g after reading σ, then it has abandoned it for a conjecture generating the language $L_\sigma \neq L_g$, and finally it has returned to a correct conjecture for L_g. This shows that $\mathbf{M}$ is in fact forced to have a U-shaped behaviour on any class containing at least L_σ and L_g. The other ingredients of the proofs are needed to ensure learnability and will be disregarded in this discussion. Now σ is a finite sequence, and L_σ is a finite language, which can be thought of as a finite table of exceptions. Instead, the language L_g containing L_σ is, at least in principle (and we have no way to decide it) an infinite set. Such an infinite set can be described, and learned, only by conjecturing a general rule.

It would be far-fetched to draw definitive conclusions about human learning from the features of the above proof. Note that the order in which finite tables and general rules come into play in the argument sketched above to enforce U-shapedness is *different* from the classical, empirical account for humans. The finite table σ is so chosen that the learner *in the first phase* commits apparent overgeneralization in the sense of conjecturing a grammar for a larger language, *in the second phase* correctly learns the finite table σ, and finally is forced to return to a conjecture for the target language containing the finite table σ and something more.

Still, the above proof exemplifies how U-shaped behaviour may be caused by the learner having to deal with two categories of objects: a finite table (of 'exceptions') *contained* in a larger language which is possibly infinite. The relations between finite and infinite members of the target class is also critical in other results discussed above. E.g., in the context of Memoryless Feedback Learning, U-shapes become *unnecessary* when classes containing *only infinite languages* are considered. Furthermore, showing that U-shapes are *unnecessary* for iterative learning of classes consisting of infinite languages turned out to be *significantly easier* than obtaining the result for arbitrary classes. In the context of Explanatory Learning, if a class does not contain an extension *of every finite set*, then that class can be learned by a decisive learner [15]!

It should be noted that the same interplay is at the very heart of many fundamental results in Gold-style learning theory — including unrelated to U-shaped learning — most notably the

seminal result showing that the class of regular languages is not learnable in the limit in the explanatory sense. In this sense the learnability of mixed classes of finite and infinite languages might be more widely relevant to the understanding of 'the problem of generalization' as a central problem of learning [45]. Note that these matters, in some cases, are also embedded in just infinite languages, where one may have some special finite parts and the other not, so these kinds of considerations *may* not go away in all cases if we confine ourselves to infinite languages.[22]

6 Other Forms of Non-Monotonic Learning

U-shapes are but an instance of non-monotonic learning. Other non-monotonic patterns have been experimentally documented and studied in a variety of cognitive-developmental situations [32,33].

In [15] a number of variants of non-monotonic learning criteria have been investigated in the context of Gold's model. In particular, the following restrictions on the learner's behaviour have been studied. (1) No return to previously abandoned wrong hypotheses, (2) No return to overinclusive hypotheses, (3) No return to overgeneralizing hypotheses and (4) No inverted U-shapes. In each case 'no return' means no *semantical* return to a previously abandoned conjecture of the specified kind. An *overinclusive* conjecture is a conjecture for a language that contains non-elements of the target language. An *overgeneralizing* conjecture is a conjecture for a language that properly includes the target language as a strict subset. An *inverted U-shape* means returning to a wrong conjecture while making a correct guess in-between.

A fairly complete picture has been obtained of the impact of the above constraints on Explanatory Learning, Vacillatory Learning and Behaviourally Correct Learning.

The general picture that emerged is the following. Forbidding return to previously abandoned wrong conjectures turns out to be a very restrictive requirement in all models. For explanatory and vacillatory learners, this amounts to imposing the strongest form of monotonicity, i.e., decisiveness (no return to any previously abandoned conjecture). In the case of Behaviourally Correct Learning, the requirement turns out to be incomparable with non-U-shapedness. Sometimes non-U-shaped behaviourally correct learners can do better than those not returning to wrong conjectures, but sometimes it's the other way round. The results about the other forms of non-monotonic learning confirm the extreme sensitivity of the Vacillation Hierarchy to this kind of constraint. Forbidding return to overinclusive hypotheses or forbidding inverted U-shapes causes the hierarchy to collapse to plain Explanatory learning, just as in Theorem 4 for U-shaped learning. Forbidding return to overgeneralizing hypotheses does not cause collapse but does restrict learning power at each level of the hierarchy. By contrast, the variants (2), (3) and (4) are useless for Explanatory and Behaviourally Correct Learning!

It is an interesting direction for future research to investigate the impact of the above variants of non-monotonic learning in models with memory limitations.

7 Discussion and Conclusion

Gold's model has been widely discussed in the Cognitive Science literature. Most commentators, however, have focused on a fundamental negative result of Gold's [40] rather than on the model as a whole and its extensions in subsequent research [50]. The theorem in question (usually

[22] And it has long been argued (except by finitists) that each natural language is infinite.

referred to as Gold's Theorem *tout court*) shows that no superfinite language class is learnable in the limit from positive data.[23]

Gold's Theorem is often invoked as evidence for the nativist approach to cognition (for recent examples see the influential [44,68]). If not to show that Universal Grammar is a "logical necessity" [68], Gold's Theorem is sometimes invoked in a less drastic way as indicating that domain-general knowledge is impossible, and that constraints on the learning process are necessary for learning — be they innate or acquired (see, e.g., [93]). As nicely observed and documented in [45], commentators usually disregard the fact that Gold himself proved in his original paper [40] that constraints on the learning process do enhance learning power (the result is that the class of c.e. languages is learnable in the limit from positive data if the learner is only required to converge on primitive recursive texts).[24] In this respect inductive learning in the limit is on a par with connectionist, statistical or Bayesian models of learning and cognition: some language classes become learnable at the cost of extra assumptions on the input or hypothesis spaces. The plausibility of the extra assumptions made in models other than Gold's can itself be questioned. As Heinz [45] puts it,

> With respect to the claim that identification in the limit makes unrealistic assumptions, I believe it is fair to debate the assumptions underlying any learning framework. However, the arguments put forward by the authors below are not convincing, usually because they say very little about what the problematic assumptions are and how their proposed framework overcomes them without introducing unrealistic assumptions of their own.

The recent debate between connectionist/emergentist and structured probabilistic inference models offers a good example of advocates of competing models mutually accusing each other of making unrealistic assumptions.[25]

The connection with linguistic nativism and the comparison with statistical models has certainly contributed to exacerbate the critiques to Gold's model from adversaries of the nativist tradition. It is not necessary to downplay Gold's model as a whole in order to defeat the argument that goes from Gold's Theorem to nativism. It is enough to observe (as both [35] and [45] do) that the class of natural languages need not be a superfinite class. This is in addition to the fact that — as nowadays widely recognized — natural languages need not coincide with a class of the Chomsky Hierarchy. Early results [1,2] show that – even if regular languages are not learnable in Gold's model – interesting and rich classes that run orthogonal to classes in the Chomsky Hierarchy are so learnable. Recent research has shown that many more interesting and rich classes orthogonal to the Chomsky Hierarchy classes *and* providing some *natural language patterns* are learnable in the limit from positive data, *and, in some cases, even efficiently so* [7,69,34,6,98].

Even advocates of the nativist tradition sometimes emphasize the limits of Gold's model, mostly in favour of statistical models of learning. Heinz [45] presents a very detailed analysis of the most common critiques to Gold's model and a convincing rebuttal of most of them. The following list of "somewhat problematic assumptions" of Gold's model can be found in the influential [68], which favours probabilistic extensions of Gold's model. (1) the learner has to identify

[23] A *superfinite* class of languages is a language class that contains some infinite language and all its finite sublanguages. Since the class of regular languages is superfinite, Gold's Theorem implies that regular languages are not learnable, and that the same is true of all larger classes of the Chomsky Hierarchy.

[24] A similarly proved positive result re a related kind of stochastic learning is found in [3].

[25] E.g., McClelland et al. [61] write, "We view the entities that serve as the basis for structured probabilistic approaches as abstractions that are occasionally useful but often misleading; they have no real basis in the actual processes that give rise to linguistic and cognitive abilities or to the development of these abilities," to which Griffiths et al. [41] reply "By contrast, we believe that greater danger lies in committing to the particular incorrect low-level mechanisms — a real possibility because most connectionist networks are vastly oversimplified when compared with actual neurons."

the target language exactly, (2) the learner receives only positive examples, (3) the learner has access to an arbitrarily large number of examples, and (4) the learner is not limited by any consideration of computational complexity. As these points are not all explicitly addressed in [45], we briefly comment on them. Note that points (1-3) apply to a family of learning criteria within Gold's setting and not to Gold-style learning theory as a whole. Concerning (1) we point out that forms of approximate learning can be and have been investigated within Gold's model [30,26,31,18]. Concerning (2) we remark that the necessity of negative information for language learning is a controversial and debated issue [11,59,92,36]. Interestingly, the fact that some form of possibly implicit negative information could be available to children learning languages was also suggested by Gold himself [40] commenting on his results on unlearnability from positive data only. For attempts at modeling partial negative information in Gold's framework see [66,4,48]. Point (3) is addressed by models with bounded memory as those mentioned in Section 4 of the present paper. The issue of computational complexity (point (4)) is in general a serious one. The difficulty of imposing a *fair* feasibility restriction on the computational complexity of the learners is indeed a serious drawback of the model. See [22] where a rigorous investigation of how some proposed solutions fail to solve the fairness problem.[26] On the other hand, *fair* and *feasible* algorithms *are known* for interesting classes orthogonal to the Chomsky Hierarchy and providing some natural language patterns (see, e.g., [69,34,6,98]).[27] The investigation of the important open mathematical questions, about the necessity of U-shapes, mentioned in prior sections, *but* where the learner is required to be both fair and feasible, has barely begun. For the humanly important memory-bounded cases, the fairness problem is apparently only difficult to sort out for the case of non-zero bounded feedback and recall queries.

To some of the points further above Clark and Lappin [35] add the requirement of convergence on all texts as an unrealistic assumption of Gold's model.[28] Oddly enough, they motivate their claim by referring to the case of feral children, who suffer from an "Impairment of learning due to an absence of data." But no text in Gold's model — however adversarial — features absence of data.[29] One might also argue that the excessively restrictive (convergence on all texts) and the excessively liberal aspects of the model (no complexity bounds) tend to even-out to some extent. As noted in [45], the restrictive aspects of Gold's learning criteria make the known *positive* results stronger rather than weaker.

One the most prominent (and usually disregarded) shortcomings of Gold's model is in our opinion its inability to model semantic information. Empirical evidence suggesting that semantics (denotation and social reinforcers) in addition to positive information might be crucial for language acquisition is presented re denotation in [64,65]. The recent and influential [63] also makes a strong case about the critical importance of social reinforcers.

Overall, we believe that some of the drawbacks of Gold's model are not unique to it and that those that are are compensated by other benefits of the model. A trade-off between applicability and predictive power on the one hand, and generality and categorical rigour on the other has to be expected at the present time. In particular — and critically for the topic of the present

[26] The *fairness* issue, first noted in [78], is the problem that, for learning in the limit, an imposition of requiring each conjecture to appear in polynomial time (in the length of the data on which it to be based) is really no useful restriction at all — since hard computations can be *un*fairly put off until a longer data sequence appears which would allow more time to compute.

[27] Considered in [20] is the possibility that human cognition is feasibly computable with massive brain parallelism providing very large coefficients for nonetheless polynomial time complexity bounds; whereas, the rest of the universe can have infeasibly computable phenomena.

[28] "Children are not generally subjected to adversarial data conditions, and if they are, learning can be seriously impaired. Therefore, there is no reason to demand learning under every presentation" [35].

[29] An adversarial text might feature absence of data up to any given finite time-bound (developmental stage) but the model has no pretension of mimicking children development to that level of detail.

paper — Gold's model is a unique setting for posing and answering with mathematical rigour questions about the *logical necessity* of learning strategies.

In general, the state of the art of Gold-style learning theory can be compared to the status of "ancient" Physics. Modeling human learning in Gold's model is weakly analogous to modeling the thermodynamics of gases without taking into account van der Waal's forces. This kind of idealization still allows some understanding of the modeled reality. Also, many other parts of Cognitive Science do not allow for precise quantitative predictions in their present form.

We thus believe that the results obtained in Gold's model can give some *insights* into human learning. As suggested in [19,45], formal results in Gold's model might guide the design of meaningful experiments. We would like to see more interaction between this branch of Computability Theory, experimental Psychology, and Cognitive Science.

As said above, the general picture that emerges from the so far known results presented in this paper is that U-shaped behaviour is unavoidable for full learning power in the context of a number of parametrized models of learning featuring a number of cognitively-motivated constraints. These results might be taken as *suggestive* of the fact that humans might exhibit U-shaped and other non-monotonic learning patterns because otherwise it would be impossible for them to learn what they need to learn to be competitive in the evolutionary marketplace. U-shaped learning could really turn out to be a "hallmark of development" [58]. Also, the results presented do illuminate from a novel perspective the critical issue of how U-shaped learning relates to the general 'problem of generalization' and to the structure of the language classes. In a number of interesting cases, the necessity of U-shaped learning disappears when learning is restricted to classes consisting of infinite languages only. These preliminary insights should be verified in two directions. (1) Empirical: by designing insightful experiments to assess measurable advantages of using a U-shaped learning strategy, and (2) Mathematical: by investigating the necessity of U-shapes in the context of more and more cognitively relevant criteria. For the second purpose, we believe that future research should focus on models of learning obtained by combining the following features: (a) vacillatory identification, (b) bounded memory, (c) queries, (d) counters, (e) fair feasibility. and also (f) stochastic elements. Re the latter, in [20] Case argues that we live in a quantum mechanical universe for which the *expected* behaviors are nonetheless algorithmic; hence, there is value in modeling the expected behavior of humans as non-stochastic but algorithmic.

We believe that each of these features is relevant for the purpose of modeling human cognition and expect that the study of U-shaped learning in models obtained by combining these features would give new insights in this interesting phenomenon.

References

1. D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980. 19
2. D. Angluin. Inference of reversible languages. *Journal of the ACM*, 29:741–765, 1982. 19
3. D. Angluin. Identifying languages from stochastic examples. 1988. Preprint. 19
4. G. Baliga, J. Case, and S. Jain. Language learning with some negative information. *Journal of Computer and System Sciences*, 51:273–285, 1995. 4, 20
5. G. Baliga, J. Case, W. Merkle, F. Stephan, and W. Wiehagen. When unlearning helps. *Information and Computation*, 206:694–709, 2008. 2, 8, 9
6. L. Becerra-Bonache, J. Case, S. Jain, and F. Stephan. Iterative learning of simple external contextual languages. *Theoretical Computer Science*, 411:2741–2756, 2010. Special Issue for *ALT'08*. 11, 19, 20
7. L. Becerra-Bonache and T. Yokomori. Learning mild context-sensitiveness: toward understanding children's language learning. In G. Paliouras and Y. Sakakibara, editors, *Grammatical Inference: Algorithms and Applications: 7th International Colloquium, ICGI 2004*, volume 3264 of *Lecture Notes in Artificial Intelligence*, pages 53–64. Springer-Verlag, October 2004. 19
8. R. Berwick. *The Acquisition of Syntactic Knowledge*. MIT Press, Cambridge, MA, 1985. 2

9. M. Bower. Starting to talk worse: Clues to language development from children's late speech errors. In S. Strauss and R. Stavy, editors, *U-Shaped Behavioral Growth*, Developmental Psychology Series. Academic Press, NY, 1982. 1, 17
10. J. Bresnan, R.M. Kaplan, S. Peters, and A. Zaenen. Cross-serial dependencies in Dutch. In W.J. Savitch, E. Bach, W. Marsh, and G. Safran-Naveh, editors, *The Formal Complexity of Natural Language*, pages 286–319. D. Reidel, Dordrecht, 1987. 4
11. R. Brown and C. Hanlon. Derivational complexity and the order of acquisition in child speech. In J. Hayes, editor, *Cognition and the Development of Language*. Wiley, 1970. 4, 20
12. S. Carey. Face perception: Anomalies of development. In S. Strauss and R. Stavy, editors, *U-Shaped Behavioral Growth*, Developmental Psychology Series. Academic Press, NY, 1982. 1
13. L. Carlucci, J. Case, S. Jain, and F. Stephan. Results on memory-limited U-shaped learning. *Information and Computation*, 205(10):1551–1573, 2007. 2, 11, 12, 13
14. L. Carlucci, J. Case, S. Jain, and F. Stephan. Non U-shaped vacillatory and team learning. *Journal of Computer and System Sciences*, 74:409–430, 2008. Special issue in memory of Carl Smith. 2, 8, 9, 10, 11, 17
15. L. Carlucci, S. Jain, E. Kinber, and F. Stephan. Variations on U-shaped learning. *Information and Computation*, 204(8):1264–1294, 2006. 2, 17, 18
16. J. Case. Periodicity in generations of automata. *Mathematical Systems Theory*, 8:15–32, 1974. 16
17. J. Case. Infinitary self-reference in learning theory. *Journal of Experimental and Theoretical Artificial Intelligence*, 6:3–16, 1994. 15, 16
18. J. Case. The power of vacillation in language learning. *SIAM Journal on Computing*, 28(6):1941–1969, 1999. 4, 5, 7, 8, 9, 10, 15, 16, 20
19. J. Case. Directions for computability theory beyond pure mathematical. In D. Gabbay, S. Goncharov, and M. Zakharyaschev, editors, *Mathematical Problems from Applied Logic II. New Logics for the XXIst Century*, International Mathematical Series, Vol. 5, pages 53–98. Springer, 2007. Invited book chapter. 21
20. J. Case. Algorithmic scientific inference: Within our computable expected reality. *International Journal of Unconventional Computing*, 2011. Journal expansion of an invited talk and paper at the *3rd International Workshop on Physics and Computation 2010*, to appear. 20, 21
21. J. Case, S. Jain, S. Lange, and T. Zeugmann. Incremental concept learning for bounded data mining. *Information and Computation*, 152:74–110, 1999. 11, 12, 13, 15
22. J. Case and T. Kötzing. Difficulties in forcing fairness of polynomial time inductive inference. In *20th International Conference on Algorithmic Learning Theory (ALT'09)*, volume 5809 of *Lecture Notes in Artificial Intelligence*, pages 263–277, 2009. 20
23. J. Case and T. Kötzing. Solutions to open questions for non-U-shaped learning with memory limitations. In M. Hutter et al., editors, *21st International Conference on Algorithmic Learning Theory (ALT'10)*, volume 6331 of *Lecture Notes in Artificial Intelligence*, pages 285–299, 2010. Expanded version invited for and accepted (with slightly new title) for the associated Special Issue of *TCS*, January 2011. 2, 12, 13, 15
24. J. Case and T. Kötzing. Strongly non U-shaped learning results by general techniques. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*. Omnipress, 2010. www.colt2010.org/papers/COLT2010proceedings.pdf is the Proceedings. 2, 9, 11, 15
25. J. Case and T. Kötzing. Measuring learning complexity with criteria epitomizers. In T. Schwentick and C. Dürr, editors, *Proceedings of the 28th International Symposium on Theoretical Aspects of Computer Science (STACS 2011)*, volume 9 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 320–331, Dagstuhl, Germany, 2011. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 15, 16
26. J. Case and C. Lynes. Machine inductive inference and language identification. In M. Nielsen and E. Schmidt, editors, *Proceedings of the 9th International Colloquium on Automata, Languages and Programming*, volume 140 of *Lecture Notes in Computer Science*, pages 107–115. Springer-Verlag, Berlin, 1982. 7, 15, 20
27. J. Case and S. Moelius. Optimal language learning. In *19th International Conference on Algorithmic Learning Theory (ALT'08)*, volume 5254 of *Lecture Notes in Artificial Intelligence*, pages 419–433. Springer, 2008. 9
28. J. Case and S. Moelius. U-shaped, iterative, and iterative-with-counter learning. *Machine Learning*, 72:63–88, 2008. 2, 11, 14
29. J. Case and S. Moelius. Parallelism increases iterative learning power. *Theoretical Computer Science*, 410:1863–1875, 2009. 11
30. J. Case and C. Smith. Anomaly hierarchies of mechanized inductive inference. In *Symposium on the Theory of Computation*, pages 314–319, 1978. 15, 20
31. J. Case and C. Smith. Comparison of identification criteria for machine inductive inference. *Theoretical Computer Science*, 25:193–220, 1983. 15, 20
32. C. H. Cashon and L. B. Cohen. The construction, deconstruction and reconstruction of infant face perception. In A. Slater and O. Pascalis, editors, *The development of face processing in infancy and early childhood*, pages 55–58. NOVA Science Publishers, New York, 2003. 18
33. C. H. Cashon and L. B. Cohen. Beyond U-shaped development in infants' processing of faces: An information-processing account. *Journal of Cognition and Development*, 5(1):59–80, 2004. 18

34. A. Clark and R. Eyraud. Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research*, 8:1725–1745, August 2007. 19, 20
35. A. Clark and S. Lappin. Computational learning theory and language acquisition. In N. Kempson, R. Asher and T. Fernando, editors, *Handbook of Philosophy of Linguistics*. Elsevier, 2010. 19, 20
36. A. Clark and S. Lappin, editors. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, Oxford, 2010. 4, 20
37. H. Friedman. *Boolean Relation Theory and Incompleteness*. Manuscript, 2011. 16
38. M. Fulk, S. Jain, and D. Osherson. Open problems in Systems That Learn. *Journal of Computer and System Sciences*, 49(3):589–604, December 1994. 2, 9
39. K. Gödel. On formally undecidable propositions of Principia Mathematica and related systems I. In S. Feferman, editor, *Kurt Gödel. Collected Works. Vol. I*, pages 145–195. Oxford Univ. Press, 1986. 16
40. E. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967. 2, 5, 7, 18, 19, 20
41. T. L. Griffiths, N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Science*, 14:357–364, 2010. 19
42. P. Halmos. *Naive Set Theory*. Springer-Verlag, NY, 1987. 3
43. L. Harrington and J. Paris. A mathematical incompleteness in Peano Arithmetic. In John Barwise, editor, *Handbook of Mathematical Logic*, pages 1133–1142. North-Holland, 1977. 16
44. M. D. Hauser, N. Chomsky, and W. T. Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298:1569–1579, 2002. 19
45. J. Heinz. Computational theories of learning and developmental psycholinguistics. Under review for the *The Cambridge Handbook of Developmental Linguistics*, 2010. 18, 19, 20, 21
46. J. Heinz and W. Idsardi. Sentence and word complexity. *Science*, 333:295–297, 2011. 4
47. S. Jain and E. Kinber. Iterative learning from positive data and negative counterexamples. *Information and Computation*, 205(12):1777–1805, 2007. 11
48. S. Jain and E. Kinber. Learning languages from positive data and negative counterexamples. *Journal of Computer and System Sciences*, 74(4):431–456, 2008. Special Issue: Carl Smith memorial issue. 20
49. S. Jain, S. Lange, and S. Zilles. Towards a better understanding of iterative learning. In J. L. Balcazar, P. Long, and F. Stephan, editors, *Algorithmic Learning Theory: 17th International Conference (ALT' 2006)*, volume 4264 of *Lecture Notes in Artificial Intelligence*, pages 169–183. Springer-Verlag, 2006. 11
50. S. Jain, D. Osherson, J. Royer, and A. Sharma. *Systems that Learn: An Introduction to Learning Theory*. MIT Press, Cambridge, Mass., second edition, 1999. 2, 3, 18
51. K. Jantke. Monotonic and non-monotonic inductive inference. *New Generation Computing*, 8:349–360, 1991. 1
52. K. Johnson. Gold's Theorem and Cognitive Science. *Philosophy of Science*, 71:571–592, 2004. 16
53. P. Johnson-Laird. *The Computer and the Mind: an Introduction to Cognitive Science*. Harvard University Press, 1988. 4
54. A. K. Joshi. How much context-sensitivity is required to provide reasonable structural descriptions: Tree adjoining grammars. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, pages 206–250. Cambridge University Press, New York, NY, 1985. 4
55. T. Kötzing. Iterative learning from positive data and counters. In *Algorithmic Learning Theory: 22th International Conference (ALT' 2011)*, pages 1–2, 2011. 14, 15
56. S. Lange and G. Grieser. Variants of iterative learning. *Theoretical Computer Science A*, 292:359–376, 2003. 11
57. S. Lange and T. Zeugmann. Incremental learning from positive data. *Journal of Computer and System Sciences*, 53:88–103, 1996. 12
58. S. Marcovitch and D.J. Lewcowicz. U-shaped functions: Artifact or hallmark of development? *Journal of Cognition and Development*, 5(2):113–118, 2004. 2, 21
59. G. Marcus. Negative evidence in language acquisition. *Cognition*, 46:53–85, 1993. 4, 20
60. G. Marcus, S. Pinker, M. Ullman, M. Hollander, T.J. Rosen, and F. Xu. *Overregularization in Language Acquisition*. Monographs of the Society for Research in Child Development, vol. 57, no. 4. University of Chicago Press, 1992. Includes commentary by H. Clahsen. 1, 2, 16
61. J. L. McClelland, M. M. Botvinick, D. C. Noelle, D. C. Plaut, T. T. Rogers, M. S. Seidenberg, and L. B. Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Science*, 14:348–356, 2010. 19
62. J. L. McClelland and K. Patterson. Rules or connections in past tense inflections: what does the evidence rule out? *Trends in Cognitive Science*, 6(11):465–472, 2002. 2
63. A. N. Meltzoff, P. K. Kuhl, J. Movellan, and T. J. Sejnowski. Foundations for a new science of learning. *Science*, 325:284–288, 2009. 20
64. D. Moeser and A. Bregman. The role of reference in the acquisition of a miniature artificial language. *Journal of Verbal Learning and Verbal Behavior*, 11:759–769, 1972. 20

65. D. Moeser and A. Bregman. Imagery and language acquisition. *Journal of Verbal Learning and Verbal Behavior*, 12:91–98, 1973. 20
66. T. Motoki. Inductive inference from all positive and some negative data. *Information Processing Letters*, 39(4):177–182, 1991. 20
67. E. Newport, L. Gleitman, and H. Gleitman. Mother i'd rather do it myself: Some effects and noneffects of maternal speech style. In C. Snow and C. Ferguson, editors, *Talking to children: Language input and acquisition*, pages 109–150. Cambridge University Press, 1977. 4
68. M. A. Nowak, N. L. Komarova, and P. Niyogi. Computational and evolutionary aspects of language. *Nature*, 417:611–617, 2002. 19
69. T. Oates, T. Armstrong, L. Becerra-Bonache, and M. Atamas. Inferring grammars for mildly context sensitive languages in polynomial-time. In *Grammatical Inference: Algorithms and Applications: 8th International Colloquium, ICGI 2006*, volume 4201 of *Lecture Notes in Artificial Intelligence*, pages 137–147. Springer-Verlag, September 2006. 19, 20
70. D. Osherson, M. Stob, and S. Weinstein. Ideal learning machines. *Cognitive Science*, 6:277–290, 1982. 2
71. D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*. MIT Press, Cambridge, Mass., 1986. 2, 6
72. D. Osherson and S. Weinstein. Criteria of language learning. *Information and Control*, 52:123–138, 1982. 7
73. S. Pinker. Formal models of language learning. *Cognition*, 7:217–283, 1979. 2
74. S. Pinker. *Language Learnability and Language Development*. Harvard University Press, 1984. 16
75. S. Pinker. Rules of language. *Science*, 253:530–535, 1991. 16
76. S. Pinker and A Prince. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193, 1988. 2
77. S. Pinker and M. T. Ullman. The past and future of the past tense. *Trends in Cognitive Science*, 6(11):456–463, 2002. 2
78. L. Pitt. Inductive inference, DFAs, and computational complexity. In *Analogical and Inductive Inference, Proceedings of the Second International Workshop (AII'89)*, volume 397 of *Lecture Notes in Artificial Intelligence*, pages 18–44. Springer-Verlag, Berlin, 1989. 20
79. K. Plunkett and V. Marchman. U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition*, 38(1):43–102, 1991. 2
80. G. K. Pullum and G. Gazdar. Natural languages and context-free languages. *Linguistics and Philosophy*, 4:471–504, 1982. 4
81. Z. Pylyshyn. *Computation and Cognition: Towards a Foundation for Cognitive Science*. MIT Press, 1984. 4
82. H. Rogers. Gödel numberings of partial recursive functions. *Journal of Symbolic Logic*, 23:331–341, 1958. 4
83. H. Rogers. *Theory of Recursive Functions and Effective Computability*. McGraw Hill, New York, 1967. Reprinted, MIT Press, 1987. 2, 4, 8, 15, 16
84. T. T. Rogers, D. H. Rakinson, and J. L. McClelland. U-shaped curves in development: a PDP approach. *Journal of Cognition and Development*, 5(1):137–145, 2004. 17
85. J. Royer. *A Connotational Theory of Program Structure*. Lecture Notes in Computer Science 273. Springer-Verlag, 1987. 4
86. J. Royer and J. Case. *Subrecursive Programming Systems: Complexity and Succinctness*. Research monograph in *Progress in Theoretical Computer Science*. Birkhäuser Boston, 1994. 16
87. D.E. Rumelhart and J.L. McClelland. On learning the past tenses of English verbs. In J.L. McClelland and D.E. Rumelhart, editors, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. II*, pages 216–271. MIT Press, Cambridge, MA, 1986. 2
88. S. M. Shieber. Evidence against the context-freeness of natural languages. *Linguistics and Philosophy*, 8:333–343, 1985. 4
89. R. Siegler. U-shaped interest in U-shaped development — and what it means. *Journal of Cognition and Development*, 5(1):1–10, 2004. 1
90. S. Strauss and R. Stavy, editors. *U-Shaped Behavioral Growth*. Developmental Psychology Series. Academic Press, NY, 1982. 1, 5
91. S. Strauss, R. Stavy, and N. Orpaz. The child's development of the concept of temperature, 1977. Unpublished manuscript, Tel-Aviv University. 1
92. N. A. Taatgen and J. R. Anderson. Why do children learn to say broke? A model of learning the past tense without feedback. *Cognition*, 86:123–155, 2002. 2, 4, 20
93. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 6022:1279–1285, 2011. 19
94. K. Wexler. On extensional learnability. *Cognition*, 11:89–95, 1982. 2
95. K. Wexler and P. Culicover. *Formal Principles of Language Acquisition*. MIT Press, Cambridge, Mass, 1980. 2, 11
96. R. Wiehagen. Limes-Erkennung rekursiver Funktionen durch spezielle Strategien. *Elektronische Informationverarbeitung und Kybernetik*, 12:93–99, 1976. 11

97. R. Wiehagen. A thesis in inductive inference. In P. Schmitt J. Dix, K. Jantke, editor, *Proceedings of the 1st International Workshop on Nonmonotonic and Inductive Logic (1990)*, volume 543 of *Lecture Notes in Computer Science*, pages 184–207. Springer Berlin/Heidelberg, 1991. 6
98. R. Yoshinaka. Learning multiple context-free languages with multidimensional substitutability from positive data. *Information and Computation*, 412:1821–1831, 2011. 19, 20
99. T. Zeugmann and S. Lange. A guided tour across the boundaries of learning recursive languages. In K.P. Jantke and S. Lange, editors, *Algorithmic Learning for Knowledge-Based Systems*, volume 961 of *Lecture Notes in Artificial Intelligence*, pages 190–258. Springer-Verlag, 1995. 1