frontiers
in Robotics and AI

PERSPECTIVE
published: 24 January 2017
doi: 10.3389/frobt.2016.00079

Future Directions in Machine Learning

Hal S. Greenwald and Carsten K. Oertel*

The MITRE Corporation, McLean, VA, USA

Current machine learning (ML) algorithms identify statistical regularities in complex data sets and are regularly used across a range of application domains, but they lack the robustness and generalizability associated with human learning. If ML techniques could enable computers to learn from fewer examples, transfer knowledge between tasks, and adapt to changing contexts and environments, the results would have very broad scientific and societal impacts. Increased processing and memory resources have enabled larger, more capable learning models, but there is growing recognition that even greater computing resources would not be sufficient to yield algorithms capable of learning from a few examples and generalizing beyond initial training sets. This paper presents perspectives on feature selection, representation schemes and interpretability, transfer learning, continuous learning, and learning and adaptation in time-varying contexts and environments, five key areas that are essential for advancing ML capabilities. Appropriate learning tasks that require these capabilities can demonstrate the strengths of novel ML approaches that could address these challenges.

Keywords: machine learning, human learning, transfer learning, continuous learning, interpretable representations, feature selection

OPEN ACCESS

Edited by:
Thomas Nowotny,
University of Sussex, UK

Reviewed by:
Michael Schmuker,
University of Hertfordshire, UK
Emili Balaguer-Ballester,
Bournemouth University, UK

****Correspondence:***
Hal S. Greenwald
hgreenwald@mitre.org

Specialty section:
This article was submitted to Computational Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *02 September 2016*
Accepted: *20 December 2016*
Published: *24 January 2017*

Citation:
Greenwald HS and Oertel CK (2017) Future Directions in Machine Learning.
Front. Robot. AI 3:79.
doi: 10.3389/frobt.2016.00079

INTRODUCTION

Machine learning (ML) algorithms (MLAs) have demonstrated success at detecting fraud (Bolton and Hand, 2002), extracting patterns from images and videos (Yang et al., 2015), playing strategy games like chess and Go (Silver et al., 2016), guiding autonomous vehicles (Enzweiler, 2015), and other applications requiring the identification of statistical patterns in multidimensional data sets, particularly when large, labeled training sets are readily available. However, they lack the rich capabilities associated with human learning that allow humans to generalize from small numbers of exemplars, apply previously learned knowledge to new tasks, and cope with changing contexts and dynamic environments. This perspective paper introduces the relevant issues at a high level (rather than providing an exhaustive literature review) and explores potential opportunities to make ML more humanlike.

DEFINING LEARNING

Depending on one's background and context, the term "learning" can mean "what one does in school," "what one gains from experience," or "an association between a conditioned stimulus and an unconditioned response." Here, we define learning as the process of acquiring and encoding knowledge for the purpose of recognizing trends, categorizing items or events, predicting the future state of the world and hypothesizing how one's actions might influence it, and performing novel tasks. One could consider the process of acquiring and encoding information to be separate from processes for prediction, reasoning, and hypothesis generation, but we take the perspective from ML research that learning software encompasses acquiring and storing knowledge as well as reasoning about the stored information, although different software components may perform these functions.

MACHINE LEARNING

The essential element of MLAs is that the information such algorithms use to categorize new information and make predictions is acquired through experience and exposure, which allows flexibility that cannot be achieved by following prespecified, deterministic rules. Conventional MLAs have been very successful in a range of domains including fraud detection, pattern recognition for images and speech, and the prediction of traffic patterns. However, it appears that "deep learning" and other recently developed approaches have made primarily incremental steps toward creating algorithms capable of highly flexible, generalized learning. The most significant advances in ML in recent years have resulted from the availability of both large, annotated data sets and greater computing power (Yamins and DiCarlo, 2016), which have enabled MLAs to represent more data with larger, more complex data structures. Fundamentally, ML systems map input features or stimuli to a set of known (supervised learning) or unknown (unsupervised learning) classes or responses. ML systems can be highly accurate in learning these relationships, but they often cannot explain how or why these relationships exist.

One of ML's biggest challenges is the availability of sufficient training data. Unless there are enough training examples to fully sample the underlying distributions of the classes or concepts being learned, ML tools cannot reliably recognize or associate novel data with the appropriate class or concept. In particular, some ML approaches like artificial neural networks and deep learning algorithms require large numbers of training examples to "converge" to a state in which the algorithms yield good performance. Also, an MLA's ability to generalize beyond the training examples typically depends on having a sufficiently sampled data space. This need for large, typically expensive corpora of labeled training data specific to each new task limits the application of ML tools and strongly contrasts with human learning capabilities.

HUMAN LEARNING

Despite the successes of conventional MLAs, their abilities to generalize from a few examples and generate hypotheses remain extremely limited when compared with typical human learning abilities. Humans benefit from multiple types of learning and adaptation mechanisms and can learn about relationships between diverse sets of information without always having a teacher present and without necessarily having to be exposed to large volumes of data. Bloom's taxonomy (Krathwohl, 2002) identifies five higher levels of knowledge-based learning of which humans are capable, including understanding (interpreting, describing, stating main ideas), applying knowledge (to novel situations), analyzing (identifying motives or causes, generalizing), evaluating, and creating (synthesizing). All of these levels are, for the present, beyond the abilities of even the most advanced ML systems, which are typically limited to the basic level of Bloom's taxonomy: recognizing and recalling associations, classifications, and categorizations. Human learning provides examples of the capabilities we would like machines to have, and the human brain can provide inspiration for the types of approaches that may enable such enhancements.

Humans are very capable learners; we are able to acquire information from explicit teaching, observation, and experience and apply it to many types of tasks. While MLAs often perform well on classification and estimation tasks, these only represent a subset of human learning. Other aspects include learning temporal sequences of events and their consequences and learning sequences of actions necessary to accomplish an objective. Humans are also capable of examining, understanding, and applying learned information to make inferences and generate insights from integrating seemingly disparate information. Our greatest capability is likely pattern recognition, which allows us to recognize similarities between otherwise seemingly disjoint data points and focus on the more relevant aspects of events and objects we observe. It seems likely that being able to identify similarities is largely responsible for enabling humans to learn from few examples; we have a store of background information and experiences on which to draw, and this enables us to bootstrap the learning process. This store of knowledge also enables humans to predict the outcomes of events, even though the particular combination of factors may never have previously been encountered. This knowledge is apparently stored using a robust, flexible representation that facilitates recall using semantic associations such as "looks like," "is like," and "is found in." Humans can update beliefs in response to errors and inconsistencies and can account for the influences of context and other factors when making predictions and generating hypotheses.

CHALLENGES AND OPPORTUNITIES

Feature Selection

Selecting the appropriate set of features from the input data determines how effectively MLAs can perform. If one has information-rich features with non-overlapping distributions corresponding to different classes (i.e., categories), then learning to differentiate between the classes is easy, and the choice of MLA is unimportant since any reasonable classifier should be effective. Conversely, if the selected features do not distinguish between the classes, then no MLA will be able to learn the correct classifications. There are various approaches to feature selection, but a common approach is to identify input features that are related to the outputs based on correlations or other statistical measures and to weight them according to their relative reliabilities. Principal component analysis, singular value decomposition, factor analysis, and other dimensionality reduction techniques can also reveal weighted combinations of features that discriminate optimally between output classes. Some ML systems have attempted to automate feature selection and use techniques like regularization, which biases the selection of potential learning models toward those with fewer numbers of parameters when performance is similar (Domingos, 2012).

Humans appear to have very robust feature extraction, storage, and selection capabilities. Olshausen and Field (1997) showed that the collection of receptive fields in the early visual system forms a sparse set of basis functions that resembles the outputs of independent component analysis applied to natural images. There is also a body of work showing that humans combine information both within and between sensory modalities in a statistically optimal way according to the relative reliabilities of the various sensory cues (Ernst and Banks, 2002; Knill and Saunders, 2003).

Even when appropriate features are available and have been identified, selecting the appropriate precision of the features can be critical. Suppose that patient age is an important feature for diagnosing a specific medical condition. Is it sufficient to distinguish between adults and children, or are the relevant differences measured in decades, years, or months (e.g., for childhood diseases)? It might seem intuitive to use the highest available precision, but higher precision requires more training data and can lead to overfitting; conversely, too little precision can lead to underfitting. The challenge of selecting the appropriate number and precision of features is known as the bias and variance problem (Geman et al., 1992; Domingos, 2012).

Most ML classifiers require the selection of a set of input features and a set of results or responses (e.g., classes) when the learning models are trained. If the input features change or the set of responses is modified (by adding classes, removing classes, or repartitioning existing classes), the MLAs typically need to be retrained from scratch (Polikar et al., 2001). In contrast, humans are capable of shifting to greater or lesser degrees of specificity with apparent ease as required by different tasks.

Machine learning systems are able to perform well as long as the test data are relatively close in feature space to examples provided in the training data set. However, ML systems tend to be brittle in that they do not perform well (if at all) on test data that are outside of the trained space. The exception to this lack of robustness is when certain ML systems are performing anomaly detection, where anything that is substantially different from the training data is considered anomalous (Lavin and Ahmad, 2015). Human learning performance with novel stimuli tends to be much more robust, although the degree to which this robustness can be attributed to a large volume of life experience (e.g., even a novel situation or stimulus will often have some similarity to a previous situation or stimulus) versus to an innate ability to generalize is an open question.

Robust Representation Schemes and Interpretability

Representation schemes determine how learned information is stored, what information is preserved, and, ultimately, the robustness of an MLA's capabilities (Bengio et al., 2013). The choice of representation scheme determines the ease with which other component processes can access and make inferences about the learned information. It also influences the ease of combining stored concepts and capturing details about contexts and causal relationships. What is most important about the representation scheme is not the representation itself but rather the set of functions and operations it enables (Bottou, 2013). For example, whole numbers can be represented without any loss of information using Roman numerals, but using a decimal encoding scheme makes basic mathematical operations much more straightforward. Developing a flexible representation scheme that preserves information while supporting critical cognitive operations is a key challenge for building more capable ML approaches. Certain types of representation schemes may make incorporating new information difficult. If a multidimensional data space is defined based on the statistics of an initial training set and new data arrives that is significantly different from the statistical distribution of the training examples previously encountered, the original representation scheme may no longer be appropriate for capturing meaningful variance and may be difficult to adapt without retraining. Ideally, there would be statistical methods for identifying the most informative dimensions that are sufficiently flexible to handle novel data without retraining from scratch. The critical challenge from an ML perspective is to understand the principles associated with the type(s) of representations necessary to enable such robust learning capabilities.

Two related criticisms of many MLAs are that their representations are not easily interpreted by humans and that there are no explanations or justifications for the results (predictions and classifications) they produce. These are important criteria for evaluating the soundness of human judgments and decisions, and having rationales and explanations available from MLAs is practically necessary for building confidence and creating trust. There is often a tradeoff between performance and interpretability that is tied to the complexity of an algorithm's representation scheme. At one extreme are simple representations for which it is easy to make sense of the meaning and significance of the parameters but that are less capable of capturing complexities in the data; at the other are representations that are uninterpretable but perform very well on difficult, complex problems. Recent research by Landecker (2014) and Turner (2015) treated MLAs as black boxes and made inferences about their internal representations by fitting models to their observed inputs and outputs; they then used these parameterized models to generate explanations for the observed classifications. Both authors argued that interpretable representations are unnecessary as long as one can fit a model to the black box MLA that provides the desired insights. However, this seems inefficient, especially if one assumes that the black box ML model is continuously learning, because both the ML model and the meta-ML model (i.e., the model of the black box model) require ongoing maintenance. Moreover, the meta-ML model is necessarily suboptimal due to the data processing inequality (Cover and Thomas, 2006), which states that additional processing cannot increase information content. MLAs' internal representations need not be straightforward for humans to interpret upon simple examination, but associated algorithms responsible for logical reasoning and inference (Davis and Marcus, 2015) should be able to use the learned information and causal relationships for logical reasoning and inference without also requiring access to the original training data. Interfaces to this data would enable components to be combined into larger architectures rather than requiring monolithic, end-to-end learning systems. Also, it would be very useful if MLAs could draw on their internal representations to provide insights into how they arrived at their outputs.

Transfer Learning and "One-Shot Learning"

Humans often make accurate inferences given a single example of a novel stimulus or situation. From as early as age two, humans are often able to recognize a novel object category based on seeing a single instance of the category and learn the meanings of new words upon first exposure (Bloom and Markson, 1998). However, most ML-based object categorization algorithms require large numbers of training examples directly related to the task before being able to accurately identify objects as belonging to a class or category (Fei-Fei et al., 2006; Held et al., 2016). Humans' ability to learn from a small number of examples is likely due to our ability to draw on experience, interactions, and learned information to bootstrap new learning. Our lifelong acquisition of knowledge helps us identify features that best define and differentiate categories and are most relevant to particular tasks. Similarly, leveraging previously learned information and procedures to build predictive world models can reduce the training time and exposure to large quantities of new task-related exemplars MLAs require (Fei-Fei et al., 2006), but most current MLAs still struggle to accomplish this. The challenge is not simply to build a common body of knowledge across tasks; a prerequisite is that the algorithm must recognize similarities between tasks and identify which information is relevant. Transferring too much irrelevant information between tasks can interfere with performance. An MLA would successfully demonstrate transfer learning when the required training time on a new task is significantly less when using data from a separate, previously learned task than when the model is trained from scratch. While researchers are familiar with the associated challenges (Pan and Yang, 2010; Senator, 2011), many MLAs remain limited in their abilities to transfer learned information across real-world tasks. However, multitask learning models (Caruana, 1997; Ando and Zhang, 2005; Pan and Yang, 2010) are capable of sharing training data across tasks and building common representations when learning multiple tasks simultaneously. These MLAs have demonstrated better generalization, reductions in the volume of required training data, and faster runtime performance relative to equivalent MLAs trained in isolation (Torralba et al., 2007). Caruana (1997) reported that overall training time for his multitask models was less than the time needed to train each task individually and that, although each training epoch required more computation, fewer epochs were required.

Continuous Learning

The artificial division between training and testing that is often applied to ML approaches seriously limits their capabilities compared to human learners (Hamker, 2001). Often, training occurs in batches and continues until additional training no longer improves performance on a specified testing benchmark, at which point learning explicitly halts. In contrast, human learning appears to be a continuous, asynchronous process (or processes) that continues throughout an individual's lifetime (Thrun, 1996), even in the absence of external stimulation. Ongoing cognitive processes like thought and self-reflection can lead to new ideas, hypotheses, and predictions, and humans learn from observing the outcomes of their own actions and behaviors. One justification for stopping learning is to prevent errors due to overfitting and overgeneralization and from learning invalid information. If the training samples are independent and identically distributed, then having more samples will always yield better performance, but these assumptions are not always necessarily valid. Sometimes humans make erroneous inferences that can lead to learned misbeliefs, but these mistakes are often corrected over time when misbeliefs lead to erroneous predictions or the individual encounters contrary evidence.

Learning and Adaptation in Time-Varying Contexts and Environments

Like transfer learning, adapting to changing environments and other such non-stationary problem spaces (e.g., changing location, time of day) has been a goal of ML for decades. Typical MLAs require periodic retraining to cope with changing (or slowly drifting) contexts and environments. Some approaches adapt to changes by learning new associations between features, contexts, and responses as they are encountered and dropping associations that have not been recently or frequently encountered to prune those that are incorrect or no longer valid. These approaches, however, typically require a static feature space, which limits the extent of their adaptability. In contrast, humans are not only capable of adapting to changing contexts and environments in many circumstances but are also able to incorporate new features and modify the attributes of existing features. Other techniques involving Kalman filters, for example, allow MLAs to adapt to changing circumstances when models' predictions differ from observations (Haykin, 2001). Ideally, ML systems would be able to adapt to changing contexts and environments but be able to recognize that a context or environment is the same as or similar to a previously encountered scenario and use the previously learned characteristics without significant retraining, which is something a Kalman filter cannot do.

CHALLENGE PROBLEM SPECIFICATIONS

Learning most frequently occurs in the context of tasks, which provide a purpose for learning how to perform an action or storing information about a particular concept, event, or entity. Having a specific task also helps to make the learning problem less abstract.

An appropriate learning challenge problem should

- Be sufficiently difficult but not too difficult (the Goldilocks Principle; Graesser et al., 2009).

- Have solutions that are verifiable but not obvious to establish a successful MLA's utility.
- Discourage the use of heuristics, guessing, and cheating to ensure that the problem must be solved using learning capabilities.
- Demonstrate the strengths of successful learning approaches by virtue of their ability to solve the challenge problem.

A suitable challenge problem should also focus on one or more of the strengths of human learning such as adaptability to changing contexts and/or environments, generalizing from limited training examples, or transferring learned knowledge from one problem to another. The challenge problem should discourage the use of brute force approaches and avoid requiring extensively large training sets that are rarely available in real-world situations.

Machine learning algorithms have been applied to a broad variety of domains and problems, and it is acceptable to test novel approaches using existing challenge problems that have previously been used with conventional MLAs. Novel ML approaches with human-like capabilities should demonstrate task performance at levels similar to (or higher than) those of conventional ML approaches; higher performance should not be a requirement, especially since some conventional approaches are statistically optimal. More importantly, the performance of novel MLAs should degrade less than conventional algorithms when the amount of training data is decreased, and novel MLAs should be capable of learning, performing, and appropriately transferring knowledge across multiple tasks.

One approach is to train a system on multiple exemplars of several object classes and then require it to classify objects into these classes. The system would also need to identify objects in the testing set that do not fit into the trained categories and group these objects by how they fit into additional categories defined along the same dimensions as the trained classes. Another approach is to use a problem that spans the chasm between sub-symbolic learning (e.g., pattern recognition) and symbolic learning (e.g., language learning). Humans are adept at both kinds of learning, but ML systems are generally focused on one or the other.

OUTLOOK

Machine learning has made tremendous advances in recent decades, especially with the availability of increased processing and storage capabilities, but there is the growing recognition that even with greater computing resources, current approaches will not reach the level of human learning capabilities. In November 2015, ETH Zurich (Switzerland) and the Max Planck Institute (Germany) announced the formation of a joint research center with experts in ML, perception, computer vision, and robotics that they hope will bridge the gap between biological and artificial learning systems. DARPA is also funding about a dozen of 12-month seedling studies that are investigating various aspects of learning. IARPA has a current program that is attempting to extract MLAs from the structure of neural tissue. A recent doctoral dissertation (Lake, 2014; Lake et al., 2015) compared human learning with ML and considered many of the same issues we have been exploring; specifically, it examined the composition of representations and concepts from more basic concepts, causality and generative models, and transfer learning. The dissertation contained a set of symbolic learning tasks with which to compare human and ML performance and a Bayesian computational model that uses compositionality, causality, and "learning-to-learn" to build generative models from one or a handful of examples. The author claimed that humans and computers differ most in the amount of training data they require for learning and in the breadth of tasks they can perform and criticized deep learning approaches for not broadening conceptual abilities.

Much of the current knowledge about learning in biological organisms comes from psychological experiments performed over the past century that characterized learning in the context of behavior using humans, non-human primates, rodents, and other animals. Behaviorists like Pavlov and Skinner demonstrated that dogs, rodents, and pigeons can learn associations between paired stimuli and between actions and consequences. However, the behaviorists cared little about the neural mechanisms responsible for mediating such learning. More recently, many neuroscience experiments have investigated the neural circuits involved in the sea slug (Aplysia) gill withdrawal reflex (since sea slugs have a simple nervous system with large neurons) and aspects of primate learning at the cellular and molecular levels. Brain imaging techniques, including functional connectivity measures, and lesion studies have also revealed which brain structures are most involved in learning. Translating these results into computational principles is not straightforward due to the complexity of the circuits and the roles of various chemical neuromodulators but also because it is still not yet sufficiently understood how individual neurons, neural populations, and neural circuits represent and encode information.

Defining requirements for novel ML approaches is easy; the challenge is in replacing the necessary but currently ill-defined steps with even preliminary sketches of algorithms and approaches that can facilitate progress toward satisfying the requirements. Besides the obvious differences between brains and computers, the former having evolved over millions of years using organic materials like proteins and fats and the latter having been engineered with metals, there are fundamental architectural differences that dictate their potential capabilities. Specifically, computer processes, whether running as single threads or as multiple parallel threads, are event driven and controlled by a master process that relies on separate structures for processing and memory. Brains are composed of complex neural networks that are responsible for both processing and information storage and operate in an asynchronous, oscillatory, massively parallel manner. Is it reasonable to expect a computer to offer the same capabilities as a self-organizing, complex adaptive system with emergent properties and behaviors? Ideally, breakthroughs in neuroscience will reveal how the brain encodes information and the computations that are responsible for enabling learning in biological systems so that we might know which elements of biological brains are necessary and/or sufficient for learning to occur (Floreano et al., 2014), but this knowledge is not strictly necessary for continued progress in artificial learning. Even if we do not entirely understand hippocampal learning circuits or neural information representation and encoding, we know that the brain supports learning more robustly and capably than current ML approaches and that it is theoretically possible to develop computational techniques with similar capabilities. Whether current computer architectures are sufficient to support transfer learning, robust data representations, and the other concepts we have discussed remains a central issue.

The challenges we have highlighted in this perspective article have been discussed before, sometimes perhaps in greater detail. However, consolidating these key points in a single article emphasizes that despite their success across various domains, MLAs are still largely special-purpose tools that lack the robustness and generalizability observed with human learning. While there have been prior attempts to build prototype MLAs that exhibit individual properties like lifelong learning and robust representation schemes, we have yet to see MLAs that offer convincing human-like learning capabilities at scale. Without these, MLAs will continue to grow as large and complex as available computing resources allow, but they will always depend on special-purpose training that will be difficult to maintain in the face of changing tasks and contexts. We hope that this paper will encourage other ML researchers to leverage insights from human learning capabilities and address challenge problems that demonstrate the advantages of bio-inspired approaches. The resulting advances in computing and autonomous systems would be expected to yield profound scientific and societal impacts.

AUTHOR CONTRIBUTIONS

Both the authors contributed equally to this work.

ACKNOWLEDGMENTS

The motivation for this work grew from discussions at the 2015 Neuro-Inspired Computational Elements (NICE) Workshop co-sponsored by DARPA, IARPA, and DOE. The authors would like to acknowledge helpful feedback and insights from Rod Holland, Reggie Carey, Brian Colder, Tod Levitt, Lashon Booker, and their reviewers.

FUNDING

This work was jointly funded by the MITRE Innovation Program's Information Technology and Technology Futures Investment Areas under the MITRE-sponsored research project "Learning Study."

REFERENCES

Ando, R. K., and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* 6, 1817–1853.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi:10.1109/TPAMI.2013.50

Bloom, P., and Markson, L. (1998). Capacities underlying word learning. *Trends Cogn. Sci.* 2, 67–73. doi:10.1016/S1364-6613(98)01121-8

Bolton, R. J., and Hand, D. J. (2002). Statistical fraud detection: a review. *Stat. Sci.* 17, 235–255. doi:10.1214/ss/1042727940

Bottou, L. (2013). From machine learning to machine reasoning. *Mach. Learn.* 94, 133–149. doi:10.1007/s10994-013-5335-x

Caruana, R. (1997). Multitask learning. *Mach. Learn.* 28, 41–75. doi:10.1023/A:1007379606734

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory, Second Edition*. Hoboken, NJ: John Wiley & Sons.

Davis, E., and Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58, 92–103. doi:10.1145/2701413

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. doi:10.1145/2347736.2347755

Enzweiler, M. (2015). The mobile revolution – machine intelligence for autonomous vehicles. *Inform. Technol.* 57, 199–202. doi:10.1515/itit-2015-0009

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi:10.1038/415429a

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Learn.* 28, 594–611. doi:10.1109/TPAMI.2006.79

Floreano, D., Ijspeert, A. J., and Schaal, S. (2014). Robotics and neuroscience. *Curr. Biol.* 24, R910–R920. doi:10.1016/j.cub.2014.07.058

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* 4, 1–58. doi:10.1162/neco.1992.4.1.1

Graesser, A., Chipman, P., Leeming, F., and Biedenbach, S. (2009). "Deep learning and emotion in serious games," in *Serious Games: Mechanisms and Effects*, eds U. Ritterfeld, M. Cody, and P. Vorderer (New York: Routledge), 81–100.

Hamker, F. H. (2001). Life-long learning cell structure – continuously learning without catastrophic interference. *Neural Netw.* 14, 551–573. doi:10.1016/S0893-6080(01)00018-1

Haykin, S. (2001). *Kalman Filtering and Neural Networks*. New York: John Wiley & Sons.

Held, D., Thrun, S., and Savarese, S. (2016). Robust single-view instance recognition. *IEEE Int. Conf. Robot. Automation* 2152–2159. doi:10.1109/ICRA.2016.7487365

Knill, D. C., and Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Res.* 43, 2539–2558. doi:10.1016/S0042-6989(03)00458-9

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory Into Pract.* 41, 212–218. doi:10.1207/s15430421tip4104_2

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science* 350, 1332–1338. doi:10.1126/science.aab3050

Lake, B. M. (2014). *Towards More Human-Like Concept Learning in Machines: Compositionality, Causality, and Learning-to-Learn*. Doctoral dissertation, Cambridge, MA: Massachusetts Institute of Technology.

Landecker, W. (2014). *Interpretable Machine Learning and Sparse Coding for Computer Vision*. Doctoral dissertation, Portland, OR: Portland State University.

Lavin, A., and Ahmad, S. (2015). "Evaluating real-time anomaly detection algorithms – the Numenta Anomaly Benchmark," in *IEEE 14th International Conference on Machine Learning and Applications*, 38–44.

Olshausen, B. A., and Field, D. J. (1997). Sparse coding with an over complete basis set: a strategy employed by V1? *Vision Res.* 37, 3311–3325. doi:10.1016/S0042-6989(97)00169-7

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowledge Data Eng.* 22, 1345–1359. doi:10.1109/TKDE.2009.191

Polikar, R., Udpa, L., Udpa, S. S., and Honavar, V. (2001). Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Trans. Syst. Man Cybernetics* 31, 497–508. doi:10.1109/5326.983933

Senator, T. E. (2011). Transfer learning progress and potential. *AI Magazine* 32, 84–86.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi:10.1038/nature16961

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? *Adv. Neural Inform. Process. Syst.* 8, 640–646.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 854–869. doi:10.1109/TPAMI.2007.1055

Turner, R. (2015). "A model explanation system," in *Black Box Learning and Inference, Neural Information Processing Systems Workshop*. Montreal. Available at: http://www.blackboxworkshop.org/pdf/Turner2015_MES.pdf

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi:10.1038/nn.4244

Yang, Y., Li, Y., Fermüller, C., and Aloimonos, Y. (2015). "Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, TX, 3686–3692. Available at: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9286/9673

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.