Accepted for publication in Current Opinion in Neurobiology

What does the mind learn? A comparison of human and machine learning representations

Jake Spicer[1], Adam N. Sanborn

Department of Psychology, University of Warwick, Coventry, UK

Abstract

We present a brief review of modern machine learning techniques and their use in models of human mental representations, detailing three notable branches: spatial methods, logical methods and artificial neural networks. Each of these branches contain an extensive set of systems, and demonstrate accurate emulations of human learning of categories, concepts and language, despite substantial differences in operation. We suggest that continued applications will allow cognitive researchers the ability to model the complex real-world problems where machine learning has recently been successful, providing more complete behavioural descriptions. This will however also require careful consideration of appropriate algorithmic constraints alongside these methods in order to find a combination which captures both the strengths and weaknesses of human cognition.

Keywords: Machine Learning, Behavioural Modelling

Introduction

A common question in the field of cognitive science: how does one go about organising all the many items and experiences encountered in everyday life into a form which is both usable and useful? Despite the extensive variety of real world events, people display a remarkable ability to acquire complex representational forms such as item taxonomies, latent patterns and causal structures simply through experience. Many theorists have therefore sought insight into this process using machine learning techniques, drawing on an extensive set of computational methods by which such structures could be generated in artificial agents. The nature and form of these methods, however, can vary wildly, leading to substantial differences in the generated representation, and so a wide range of behavioural predictions.

In this paper, we present a brief review of a number of these methods, considering differences in form and operation, applications to real-world phenomena, and variations in complexity. As machine learning processes encompass a wide range of systems and methodologies, providing more potential methods than can be adequately summarised within the scope of this paper, we here focus on three key branches of these systems: spatial methods, logical methods and artificial neural networks, detailing the more prominent methods within these branches in ascending complexity. While we attempt to provide brief summaries of these methods, we refer the reader to

[1] Corresponding author

Email addresses: j.spicer@warwick.ac.uk (Jake Spicer), a.n.sanborn@warwick.ac.uk (Adam Sanborn)

papers within each branch that provide greater detail on specific methods for further reading. This review primarily focuses on applications to human categorisation due to the extensive investigation of mental representations in this area, though we also present applications to other tasks where appropriate.

Spatial Methods

We define spatial methods as mechanisms which directly organise actual items and experiences, often placing these items in a multidimensional representational space. One simple form this method can take is to store all experienced items in a representational space with a similarity gradient around each for use in future prediction or classification. This is exemplified in recent kernel methods, which use a variety of similarity metrics and often remove redundant stored items to provide a more efficient representation [1]; for example, support vector machines use kernel functions to draw decision boundaries between categories [2], often resulting in highly accurate classification performance [3–5].

Such methods are mirrored in behavioural models by exemplar representations, which similarly store all items in a multidimensional feature space, using assessments of similarity to these stored items when making new predictions [1,6]. These models are therefore commonly used as a simple representation of item memory (e.g. [7,8]), as well as a base for more complex learning models (e.g. ALCOVE [9]), though concerns have been raised regarding the psychological plausibility of exemplar representations [10].

Beyond this direct representation of items, spatial methods can also produce a variety of more abstract representations, with some of the more simplistic being those that aggregate sets of items into collected averages, such as k-means clustering and self-organised maps [11,12]. These methods correspond with the use of prototypes in human learning, often contrasted with exemplar formats, which also use aggregates to represent a set, though this usually involves only a singular average [13], matching with the most basic form of these methods. Prototypes provide an intuitive method of summarising data sets into an easy-to-use form, but in doing so can miss more complex aspects of item representation, including the relations between stimuli (including the exclusive-or problem illustrated in Figure 1) [10,14], suggesting greater complexity is in fact required.

A more flexible representation is provided by clustering methods, in which items within a set are assigned to subgroups called clusters according to observed similarities. While this can involve a fixed number of clusters as in the above k-means, much machine learning research has investigated non-parametric forms of this representation, in which the number of clusters is flexible and learned from the data. The Dirichlet Process Mixture Model is one of the most notable non-parametric clustering methods, in which cluster assignments are made sequentially according to the existing number of cluster members [15]. More recently, the Indian Buffet Process prior has been used in similar systems to provide alternative representations in which cluster assignments are replaced with feature inferences, being potentially infinite in number [16].

In cognitive science, these processes have been most commonly used within Bayesian models of cognition, providing a non-parametric probabilistic system which infers external structures according to both direct observation and prior beliefs [17,18]. These clustering models essentially offer an interpolation between the above exemplar and prototype forms, with each cluster acting as a distinct prototype; any created partition therefore falls between these two extremes depending on the number of clusters formed. This is most evident in rational models of categorisation [17], though

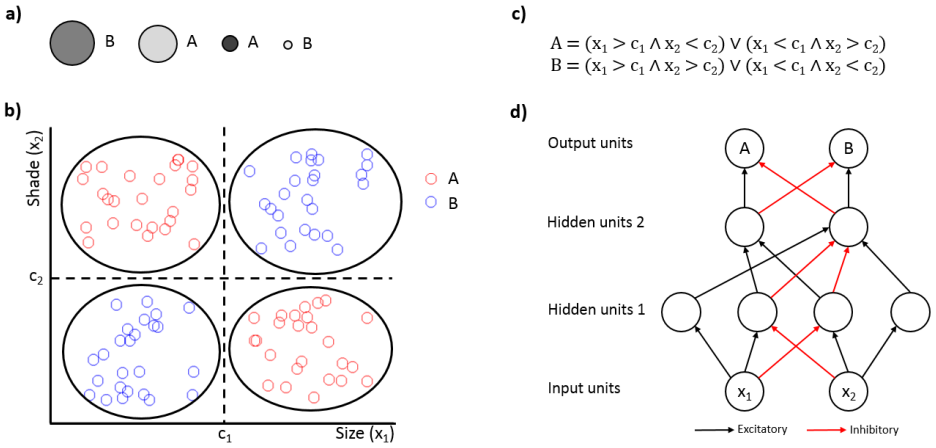

Figure 1. Applications of three machine learning methods using different representations to a basic exclusive-or classification task using the dimensions of size and shade (examples of stimuli in a). This problem can be solved by a spatial clustering method (b), a logical Boolean method (c), and an artificial neural network (d).

similar techniques have been applied to numerosity [19,20], language segmentation [21] and causal inference [22].

Recent advancements in these clustering methods have led to increasingly complex representations, including hierarchical systems where clusters can be nested within each other [23], and the CrossCat model, where multiple partitions of the same items can be formed from different feature patterns [24]. This has led to the development of hierarchical models of human categorisation [25,26], as well as structural form models which select not just the organisation of items but the form of that organisation, considering clusters, trees and chains among others [27,28]. This provides a substantial level of flexibility in the ultimate representation, but by expanding the number of considered forms in this way, such systems require strong inductive priors to adequately limit the hypothesis space in order to allow efficient learning from limited data.

Logical Methods

Logical methods define items or concepts using logical statements concerning the features of the target, identifying common elements within a data set that can be distilled into grammatical terms for use in future predictions. This is most clearly demonstrated in inductive logic programming systems [29], which have been used to generate logical rules for classifications [30] and state transitions [31].

Similar concepts can be observed in rule-based models of human behaviour, commonly used in categorisation as category membership is often defined by similar boundaries in everyday life [32,33]; indeed, Feldman suggests that categorisation behaviour reflects the use of Boolean logic, with the difficulty in learning a rule being proportional to its Boolean complexity [34]. Models such as RULEX [35] therefore search through stimulus dimensions to find the simplest rule which maximises discriminability, whilst also creating a store of exceptions. Much like the spatial methods above, these have more recently been developed into more advanced probabilistic grammars (e.g. [36]), allowing for stronger inductive inferences from limited data. This can again lead to a rational model of structure discovery, in which a rule is inferred from observations using priors on individual components to provide a bias toward simplicity, with lower probabilities for more complex, multidimensional rules. Rule-based systems are not purely limited to categorisation, however, with similar methods being applied to the learning of language [37] and functions [38].

The key advantage of these logical systems is compositionality: individual elements can be combined to create much more complex rules from fairly simplistic building blocks. In addition, grammars provide a modality-independent representation, able to be translated into alternate formats to direct behaviour in tasks beyond those used for initial learning [39]. Logical systems do, however, naturally draw hard boundaries between categories, making it more difficult to account for the graded nature of human category representations [40]. While probabilistic versions of these systems do help to account for this issue [36,41], such additional flexibility again requires strong inductive priors in order to learn effectively from limited data.

Recent years have also offered a more advanced form of this representation in 'program' models [42,43], which use Bayesian induction to construct complex production procedures from more basic elements. This is suggested to generate broad and rich representations from small data samples, allowing for accurate generalisations from even a single category member [43] and more intuitive and predictable laws in function learning [44]. Programs do, however, present an especially complex representational form, and as such are more critically in need of adequate biases to match human learning.

Artificial Neural Networks

Artificial neural networks provide an alternate form of representation using networks of interconnected nodes, with the strength of the connections being adjusted with experience to reproduce external patterns. This representation intuitively provides a closer correspondence between method principles and actual implementation in the brain: connectionist networks offer a simplified emulation of true neural structures, inherently affording such methods a degree of external validity [45]. These systems therefore contrast with both spatial and logical models of human cognition in their level of explanation; while the above methods focus on Marr's computational level, network methods instead operate at Marr's implementation level [46].

Rather than the strict delineations between methods seen in the above branches, complexity within these networks increases somewhat gradually according to size, both in terms of breadth and depth. This extends from basic mechanisms like perceptrons, which essentially provide a connectionist implementation of prototypes [1,47], to more complex parallel distributed processing systems, expanding the number of nodes and connections to create a more extensive network with a greater representational capacity [48]. There are, however, additional complexities in these methods beyond network size, with recurrent and convolutional networks being some of the more notable forms. In addition, recent neural networks have been further expanded to include external memory stores, using these elements to further improve their performance [49].

Highly complex neural networks have in fact become increasingly common in machine learning in recent years due to a surge in the use of deep learning systems in various complex tasks; these methods use multiple, hierarchical layers of connections for increasing levels of abstraction [50,51]. Such systems have the advantage of flexibility, providing a single, global system that can be applied fairly readily to multiple fields. Deep learning systems have therefore been successful in finding categorical structures in image recognition [52,53] and speech processing [54,55], as well as matching or exceeding human performance on complex tasks such as playing video and board games [56,57].

Within cognitive modelling, simple network models have been commonly used in associative learning theories (e.g. [58]), providing an extensive literature using networks often limited to only a few nodes representing basic stimulus features. The

more complex networks used in deep learning, meanwhile, are still beginning to be applied to behaviour [59-61], creating cognitive models that can take advantage of the power of such methods. There are, however, concerns whether such applications are truly valid: while deep learning systems demonstrate a similar level of performance to human learning, and use similar representations to those of actual neural systems [62,63], both speed of learning and ease of generalisation are much higher in people than machines [43,64], potentially indicating some difference in operation. This is further complicated by the opaqueness of such methods, with any generated representation being distributed across a potentially enormous series of connection weights; this can make interpretation of the learned representation difficult, relying more on behavioural predictions than any obvious structure.

Conclusions

As the above sections hopefully illustrate, machine learning methods have become increasingly common within cognitive science as descriptions of human behaviour, providing accurate emulations in various tasks. One question that could then be raised when comparing human and machine learning is which of these models is most accurate to the human learning system given their differences in representation. Such a contrast may not, however, be entirely helpful given that, as alluded to by George Box [65], no model provides a perfect description of behaviour, instead offering useful explorations of the ways in which human learning operates. As such, the value of these methods for cognitive science depends not just on their match to human behaviour, but also the suitability of the representation used to the needs and aims of the topic at hand; for example, neural networks may be best suited to subjects where the representation is less vital than the resulting behaviour, while spatial models can be used when this representation is under examination. These aspects must therefore be considered alongside evaluations of accuracy when selecting research models.

A more general aspect raised in this comparison is the difference between the goals of machine learning and cognitive science, which in turn leads to differences in approach: machine learning methods have primarily focused on efficacy, often searching for optimal real-world performance; this has led these systems to approach increasingly complex real-world tasks over the course of their development, building on their previous success in more basic problems. Conversely, cognitive science has instead sought to capture human behaviour, whether optimal or not; as such, much of the research in this field has involved simplified diagnostic versions of real-world tasks, using abstracted stimuli and designs for greater experimental control.

This contrast may then provide a direction for future cognitive research: while the existing approach is certainly highly valuable in defining the operations of human cognition, in order to provide more complete models of behaviour, cognitive scientists should also attempt to address the complex real-world tasks currently targeted by machine learning. Such an expansion can of course take advantage of the success of machine learning methods in these tasks, continuing to use these systems as a base for more advanced models of behaviour.

There is, however, another aspect that must be considered when applying these methods to human behaviour: the constraints placed upon the system to make learning feasible. In the case of machine learning, this relates to the algorithms used to define the parameters of the generated representation, determining the learnability of the system; for example, the success of the previously noted deep learning systems is not simply due to the use of complex networks, but also the advances in algorithms which make that representation learnable.

This means that any application of these advanced machine learning methods to behaviour requires careful consideration of both the generated representation and the associated algorithms used to acquire that representation. This is particularly important given that this algorithm must account for both the general level of human performance as well as any systematic errors made by human learners that would be undesirable in artificial agents. There are, however, multiple potential algorithms that could be applied to such models to fill this role according to the form of the representation; for example, sampling procedures can be used as an approximation for a number of Bayesian spatial models [66], while network structures can use prediction errors to facilitate learning [67], with both offering potential explanations for noted human biases [68-73].

We therefore conclude by advising researchers to consider a broad range of both representations and algorithms when attempting to model human behaviour, looking for the combination of these elements which best captures both the strengths and weaknesses of human learning; this will allow cognitive science to take full advantage of the power of machine learning methods, and so a greater ability to understand how people solve the complex problems found in everyday life.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest: none.

References

[1] Jäkel F, Schölkopf B, Wichmann FA: **A tutorial on kernel methods for categorization**. *J Math Psychol* 2007, **51**:343–358. doi:10.1016/j.jmp.2007.06.002.

[2] Cristianini N, Schölkopf B: **Support vector machines and kernel methods: The new generation of learning machines**. *Artif Intell Mag* 2002, **23**:31–42. doi:10.1609/aimag.v23i3.1655.

[3] Decoste D, Schölkopf B: **Training invariant support vector machines**. *Mach Learn* 2002, **46**:161–190. doi:10.1023/A:1012454411458.

[4] Razzaghi T, Roderick O, Safro I, Marko N: **Multilevel weighted support vector machine for classification on healthcare data with missing values**. *PLoS One* 2016, **11**:e0155119. doi:10.1371/journal.pone.0155119.

[5] Rasmussen M, Rieger J, Webster KN: **Approximation of reachable sets using optimal control and support vector machines**. *J Comput Appl Math* 2017, **311**:68–83. doi:10.1016/j.cam.2016.06.015.

[6] Nosofsky RM: **Attention, similarity, and the identification-categorization relationship**. *J Exp Psychol Gen* 1986, **115**:39–57. doi:10.1037/0096-3445.115.1.39.

[7] Brown GDA, Neath I, Chater N: **A temporal ratio model of memory**. *Psychol Rev* 2007, **114**:539–576. doi:10.1037/0033-295X.114.3.539.

[8] Nosofsky RM, Sanders CA, McDaniel MA: **Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain**. *J Exp Psychol Gen* 2018, **147**:328–353. doi:10.1037/xge0000369.

[9] Kruschke JK: **ALCOVE: An exemplar-based connectionist model of category learning**. *Psychol Rev* 1992, **99**:22–44. doi:10.1037/0033295X.99.1.22.

[10] Vanpaemel W, Storms G: **In search of abstraction: The varying abstraction model of categorization**. *Psychon Bull Rev* 2008, **15**:732–749.

doi:10.3758/PBR.15.4.732.

[11] Kohonen T: **Essentials of the self-organizing map**. *Neural Networks* 2013, **37**:52–65. doi:10.1016/j.neunet.2012.09.018.

[12] Biehl M, Hammer B, Villmann T: **Prototype-based models in machine learning**. *Wiley Interdiscip Rev Cogn Sci* 2016, **7**:92–111. doi:10.1002/wcs.1378.

[13] Reed SK: **Pattern recognition and categorization**. *Cogn Psychol* 1972, **3**:382–407. doi:10.1016/0010-0285(72)90014-X.

[14] Nosofsky RM: **Exemplars, prototypes, and similarity rules**. In *Essays in honor of William K. Estes: Vol. 1. From learning theory to connectionist theory*. Edited by Healy AF, Kosslyn SM, Shiffrin RM. Erlbaum; 1992:149–167.

[15] Antoniak CE: **Mixtures of Dirichlet Processes with applications to Bayesian nonparametric problems**. *Ann Stat* 1974, **2**:1152–1174. doi:10.1214/aos/1176342871.

[16] Griffiths TL, Ghahramani Z: **The Indian Buffet Process: An introduction and review**. *J Mach Learn Res* 2011, **12**:1185–1224.

[17] Anderson JR: **The adaptive nature of human categorization**. *Psychol Rev* 1991, **98**:409–429. doi:10.1037/0033-295X.98.3.409.

[18] Austerweil JL, Griffiths TL: **A nonparametric Bayesian framework for constructing flexible feature representations**. *Psychol Rev* 2013, **120**:817–851. doi:10.1037/a0034194.

[19] Gershman SJ, Niv Y: **Perceptual estimation obeys Occam's razor**. *Front Psychol* 2013, **4**:623. doi:10.3389/fpsyg.2013.00623.

[20] Sanborn AN, Beierholm UR: **Fast and accurate learning when making discrete numerical estimates**. *PLoS Comput Biol* 2016, **12**:e1004859. doi:10.1371/journal.pcbi.1004859.

[21] Goldwater S, Griffiths TL, Johnson M: **A Bayesian framework for word segmentation: Exploring the effects of context**. *Cognition* 2009, **112**:21–54. doi:10.1016/j.cognition.2009.03.008.

[22] Buchsbaum D, Griffiths TL, Plunkett D, Gopnik A, Baldwin D: **Inferring action structure and causal relationships in continuous sequences of human action**. *Cogn Psychol* 2015, **76**:30–77. doi:10.1016/j.cogpsych.2014.10.001.

[23] Blei DM, Griffiths TL, Jordan MI: **The nested Chinese Restaurant Process and Bayesian nonparametric inference of topic hierarchies**. *J ACM* 2010, **57**:7:1--7:30. doi:10.1145/1667053.1667056.

[24] Mansinghka V, Shafto P, Jonas E, Petschulat C, Gasner M, Tenenbaum JB: **CrossCat: A fully Bayesian nonparametric method for analyzing heterogeneous, high dimensional data**. *J Mach Learn Res* 2016, **17**:138:1--49.

[25] Griffiths TL, Canini KR, Sanborn AN, Navarro D: **Unifying rational models of categorization via the hierarchical Dirichlet Process**. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Edited by Sun R, Miyake N. Cognitive Science Society; 2007.

[26] Heller K, Sanborn AN, Chater N: **Hierarchical learning of dimensional biases in human categorization**. In *Advances in Neural Information Processing Systems 22*. Edited by Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A. MIT Press; 2009.

[27] Kemp C, Tenenbaum JB: **The discovery of structural form**. *Proc Natl Acad Sci* 2008, **105**:10687–10692. doi:10.1073/pnas.0802631105.

[28]** Lake BM, Lawrence ND, Tenenbaum JB: **The emergence of organizing structure in conceptual representation**. *Cogn Sci* 2018, **42**:online pre-print. doi:10.1111/cogs.12580.

An advanced spatial model that avoids using predefined structural forms, relying instead on a simplicity bias towards sparsity using Bayesian induction.

[29] Muggleton S, De Raedt L, Poole D, Bratko I, Flach P, Inoue K, Srinivasan A: **ILP turns 20: Biography and future challenges**. *Mach Learn* 2012, **86**:3–23. doi:10.1007/s10994-011-5259-2.

[30] Katzouris N, Artikis A, Paliouras G: **Incremental learning of event definitions with Inductive Logic Programming**. *Mach Learn* 2015, **100**:555–585. doi:10.1007/s10994-015-5512-1.

[31] Inoue K, Ribeiro T, Sakama C: **Learning from interpretation transition**. *Mach Learn* 2014, **94**:51–79. doi:10.1007/s10994-013-5353-8.

[32] Bruner SJ, Goodnow JJ, Austin GA: *The Study of Thinking*. Wiley; 1956.

[33] Shepard RN, Hovland CI, Jenkins HM: **Learning and memorization of classifications.** *Psychol Monogr Gen Appl* 1961, **75**:1–42. doi:10.1037/h0093825.

[34] Feldman J: **Minimization of Boolean complexity in human concept learning**. *Nature* 2000, **407**:630–633. doi:10.1038/35036586.

[35] Nosofsky RM, Palmeri TJ, McKinley SC: **Rule-plus-exception model of classification learning.** *Psychol Rev* 1994, **101**:53–79. doi:10.1037/0033-295X.101.1.53.

[36] Goodman NA, Tenenbaum JB, Feldman J, Griffiths TL: **A rational analysis of rule-based concept learning**. *Cogn Sci* 2008, **32**:108–154. doi:10.1080/03640210701802071.

[37] Frank MC, Tenenbaum JB: **Three ideal observer models for rule learning in simple languages**. *Cognition* 2011, **120**:360–371. doi:10.1016/j.cognition.2010.10.005.

[38] Lucas CG, Griffiths TL, Williams JJ, Kalish ML: **A rational model of function learning**. *Psychon Bull Rev* 2015, **22**:1193–1215. doi:10.3758/s13423-015-0808-5.

[39] Erdogan G, Yildirim I, Jacobs RA: **From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach**. *PLoS Comput Biol* 2015, **11**:e1004610. doi:10.1371/journal.pcbi.1004610.

[40] Rosch E: **On the internal structure of perceptual and semantic categories**. In *Cognitive Development and Acquisition of Language*. Edited by Moore TE. Academic Press; 1973.

[41] Shepard RN: **Towards a universal law of generalization for psychological science**. *Science (80-)* 1987, **237**:1317–1323. doi:10.1126/science.3629243.

[42] Ghahramani Z: **Probabilistic machine learning and artificial intelligence**. *Nature* 2015, **521**:452–459. doi:10.1038/nature14541.

[43] Lake BM, Salakhutdinov R, Tenenbaum JB: **Human-level concept learning through probabilistic program induction**. *Science (80-)* 2015, **350**:1332–1338. doi:10.1126/science.aab3050.

[44]** Schulz E, Tenenbaum JB, Duvenaud D, Speekenbrink M, Gershman SJ:

Compositional inductive biases in function learning. *Cogn Psychol* 2017, **99**:44–79. doi:10.1016/j.cogpsych.2017.11.002.

A demonstration of the power of program models and their correspondence with human behaviour in the learning of mathematical functions.

[45] McClelland JL, Botvinick MM, Noelle DC, Plaut DC, Rogers TT, Seidenberg MS, Smith LB: **Letting structure emerge: connectionist and dynamical systems approaches to cognition**. *Trends Cogn Sci* 2010, **14**:348–356. doi:10.1016/j.tics.2010.06.002.

[46] Marr D: *Vision: A computational investigation into the human representation and processing of visual information*. Freeman; 1982.

[47] Rosenblatt F: **The perceptron: A probabilistic model for information storage and organization in the brain**. *Psychol Rev* 1958, **65**:386–408. doi:10.1037/h0042519.

[48] Rumelhart DE, McClelland JL, the PDP Research Group: *Parallel Distributed Processing, volume 1: Foundations*. MIT Press; 1986.

[49] Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, Colmenarejo SG, Grefenstette E, Ramalho T, Agapiou J, et al.: **Hybrid computing using a neural network with dynamic external memory**. *Nature* 2016, **538**:471–476. doi:10.1038/nature20101.

[50] Lecun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**:436–444. doi:10.1038/nature14539.

[51] Schmidhuber J: **Deep Learning in neural networks: An overview**. *Neural Networks* 2015, **61**:85–117. doi:10.1016/j.neunet.2014.09.003.

[52] Farabet C, Couprie C, Najman L, Lecun Y: **Learning hierarchical features for scene labeling**. *IEEE Trans Pattern Anal Mach Intell* 2013, **35**:1915–1929. doi:10.1109/TPAMI.2012.231.

[53] Krizhevsky A, Sutskever I, Hinton GE: **ImageNet classification with deep convolutional neural networks**. *Commun ACM* 2017, **60**:84–90. doi:10.1145/3065386.

[54] Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, et al.: **Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups**. *IEEE Signal Process Mag* 2012, **29**:82–97. doi:10.1109/MSP.2012.2205597.

[55] Chen D, Mak B: **Multi-task learning of deep neural networks for low-resource speech recognition**. *IEEE/ACM Trans Audio, Speech, Lang Process* 2015, **23**:1–1. doi:10.1109/TASLP.2015.2422573.

[56] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al.: **Human-level control through deep reinforcement learning**. *Nature* 2015, **518**:529–533. doi:10.1038/nature14236.

[57] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al.: **Mastering the game of Go with deep neural networks and tree search**. *Nature* 2016, **529**:484–489. doi:10.1038/nature16961.

[58] Rescorla RA, Wagner AR: **A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement**. In *Classical Conditioning II*. Edited by Black AH, Prokasy WF. Appleton-Century-Crofts; 1972.

[59] Lake BM, Zaremba W, Fergus R, Gureckis TM: **Deep neural networks predict category typicality ratings for images**. In *Proceedings of the 37th Annual Cognitive Science Society*. Edited by Noelle DC, Dale R, Warlaumont AS, Yoshimi J, Matlock T, Jennings CD, Maglio PP. Cognitive Science Society; 2015:1243–1248.

[60]** Peterson J, Abbott JT, Griffiths TL: **Adapting deep network features to capture psychological representations**. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Edited by Papafragou A, Grodner D, Mirman D, Trueswell JC. Cognitive Science Society; 2016.

An application of deep learning to human learning, presenting potential methods for adapting deep neural networks to better match human behaviour in image classification.

[61]* Testolin A, Zorzi M: **Probabilistic models and generative neural networks: Towards an unified framework for modeling normal and impaired neurocognitive functions**. *Front Comput Neurosci* 2016, **10**:73. doi:10.3389/fncom.2016.00073.

A combination of network mechanisms and higher level computational goals, providing a more complete depiction of the neural foundations of cognition.

[62] Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ: **Deep neural networks rival the representation of primate IT cortex for core visual object recognition**. *PLoS Comput Biol* 2014, **10:**e1003963. doi:10.1371/journal.pcbi.1003963.

[63] Khaligh-Razavi SM, Kriegeskorte N: **Deep supervised, but not unsupervised, models may explain IT cortical representation**. *PLoS Comput Biol* 2014, **10:**e1003915. doi:10.1371/journal.pcbi.1003915.

[64]** Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ: **Building machines that learn and think like people**. *Behav Brain Sci* 2017, **40**:e253. doi:10.1017/S0140525X16001837.

A detailed summary of both the differences between machine and human learning and potential methods for integration.

[65] Box GEP, Hunter JS, Hunter WG: *Statistics for Experimenters (2nd ed.)*. Wiley-Interscience; 2005.

[66] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB: *Bayesian Data Analysis*. CRC Press; 2013.

[67] Rumelhart DE, Hinton GE, Williams RJ: **Learning representations by back-propagating errors**. *Nature* 1986, **323**:533–536. doi:10.1038/323533a0.

[68] Sanborn AN, Griffiths TL, Navarro DJ: **Rational approximations to rational models: Alternative algorithms for category learning.** *Psychol Rev* 2010, **117**:1144-1167. doi:10.1037/a0020511.

[69] Griffiths TL, Vul E, Sanborn AN: **Bridging levels of analysis for probabilistic models of cognition.** *Curr Dir Psychol Sci* 2012, **21**:263-268. doi:10.1177/0963721412447619.

[70]* Sanborn AN, Chater N: **Bayesian brains without probabilities.** *Trends Cogn Sci* 2016, **20**:883-893. doi:10.1016/j.tics.2016.10.003.

A summary of sampling algorithms for Bayesian models, and their potential ability to explain common human learning fallacies.

[71]* Dasgupta I, Schulz E, Gershman SJ: **Where do hypotheses come from?** *Cogn*

Psychol 2017, **96**:1-25. doi:10.1016/j.cogpsych.2017.05.001.

A rational process model which uses sampling algorithms to accurately capture a number of human learning phenomena.

[72] Jones M, Curran T, Mozer MC, Wilder MH: **Sequential effects in response time reveal learning mechanisms and event representations.** *Psychol Rev* 2013, **120**:628–666. doi:10.1037/a0033180.

[73] Sakamoto Y, Jones M, Love BC: **Putting the psychology back into psychological models: Mechanistic versus rational approaches**. *Mem Cogn* 2008, **36**:1057–1065. doi:10.3758/MC.36.6.1057.