Imagination Machines: A New Challenge for Artificial Intelligence

Sridhar Mahadevan
sridhar.i.mahadevan@gmail.com
College of Information and Computer Sciences
140 Governor's Drive
Amherst, MA 01003

Abstract

The aim of this paper is to propose a new overarching challenge for AI: *the design of imagination machines*. Imagination has been defined as *the capacity to mentally transcend time, place, and/or circumstance.* Much of the success of AI currently comes from a revolution in data science, specifically the use of deep learning neural networks to extract structure from data. This paper argues for the development of a new field called *imagination science*, which extends data science beyond its current realm of learning probability distributions from samples. Numerous examples are given in the paper to illustrate that human achievements in the arts, literature, poetry, and science may lie beyond the realm of data science, because they require abilities that go beyond finding correlations: for example, generating samples from a novel probability distribution different from the one given during training; causal reasoning to uncover interpretable explanations; or analogical reasoning to generalize to novel situations (e.g., imagination in art, representing alien life in a distant galaxy, understanding a story about talking animals, or inventing representations to model the large-scale structure of the universe). We describe the key challenges in automating imagination, discuss connections between ongoing research and imagination, and outline why automation of imagination provides a powerful launching pad for transforming AI.

> "Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world." – Albert Einstein

> 'Imagine there's no countries. It isn't hard to do. Nothing to kill or die for. And no religion too" – Song by John Lennon

Artificial intelligence is poised to become the "electricity" of our age (Ng 2016), transforming industries across a wide spectrum of areas, from autonomous driving to voice-activated virtual personal assistants. However, these successes of AI, powered by data science (Murphy 2013) and deep learning (Goodfellow, Bengio, and Courville 2016), may not be sufficient for AI to be capable of matching human capabilities in the long run. This paper focuses specifically on one core capability – imagination – and discusses why its automation may be fundamental to the continuing success of AI in the coming decades.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Figure 1: Jean-Michel Basquiat's untitled painting of a human skull sold recently at a New York auction for over 100 million dollars. Art is a paradigmatic example of the imaginative capacity of humans.

The Oxford Handbook of the Development of Imagination defines **imagination as the capacity to mentally transcend time, place, and/or circumstance** (Taylor 2013). Einstein prized imagination because it enabled him to pose hypothetical questions, such as "What would the world look like if I rode a beam of light", a question that led him to develop the revolutionary theory of special (and later, general) relativity. Imagination is a hallmark of counterfactual and causal reasoning (Pearl 2009). Imagination also provides the foundational basis for art (see Figure 1). Basquiat's painting illustrates what is special about imagination in art: fidelity to the original is not the objective here, but rather the striking use of colors and textures to signify an illusion.

In John Lennon's famous song "Imagine", he asks us to contemplate a world without countries, an abstraction of reality that is a hallmark of imaginative thinking. In Beethoven's Pastoral symphony, each of the five movements portrays a particular aspect of nature, from the slow movement depicting the motion of a stream to the strenuous fourth movement depicting the arrival of a storm and thunder. Imagination plays a central role in the lives of children and adults. The runaway success of the Harry Potter series shows what a gifted writer can accomplish in holding the attention of children, highlighting the crucial role that make-believe plays in the formative years of children. Wonder Woman was the smash $1 billion Hollywood hit of the year, showing once again that the world of fantasy and imagination is one sure fire way to create a money making movie.

Although imagination has attracted the attention of some researchers, the early work on this topic has been somewhat limited in scope (Alexander 2001), and more recent work has explored this topic in rather restricted situations (Pascanu et al. 2017; Elgamman et al. 2017). This brief paper summarizes several converging lines of argument as to why imagination machines constitutes a broad comprehensive research program that has the potential to transform AI in the next few decades. Imagination is one of the hallmarks of human intelligence (Asma 2017), an ability that manifests itself in children at a very young age, and prized by society in many endeavors, from art (see Figure 1) and literature to science. It represents an area largely ignored by most AI research, although tantalizing glimpses of the power of imagination are beginning to manifest themselves in different strands of current AI research, as will be discussed below.

As work by the Nobel-prize winning economist Daniel Kahneman (with his late colleague, Amos Tversky) has shown, based on many empirical studies, human decision making does not conform to the maxims of expected utility theory. Faced with a "lottery" (a decision problem with several uncertain outcomes with different payoffs), human decision making often does not result in picking choices that have the maximum expected utility. Year after year, in state after state, millions of Americans buy lottery tickets, because they can "imagine" themselves winning and becoming rich, despite the vanishingly small probability of winning. Clearly, for many humans, imagination in this case (mis)guides their actions into violating the principle of maximizing expected utility.

Figure 2: Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) can create images from samples of a fixed distribution, but imaginative art such as Basquiat's painting in Figure 1 require going beyond reproducing existing art. A variant of a GAN, called a "creative adversarial network" attempts to generate "novel" art (Elgamman et al. 2017), producing the images shown above.

From Data Science to Imagination Science

"A theory is not like an airline or bus timetable. We are not interested simply in the accuracy of its predictions. A theory also serves as a base for thinking. It helps us to understand what is going on by enabling us to organize our thoughts. Faced with a choice between a theory which predicts well but gives us little insight into how the system works and one which gives us this insight but predicts badly, I would choose the latter, and I am inclined to think that most economists would do the same." – Ronald Coase, Nobel-prize winning economist.

"I now take causal relationships to be the fundamental building blocks both of physical reality and of human understanding of that reality, and I regard probabilistic relationships as but the surface phenomena of the causal machinery that underlies and propels our understanding of the world". – Judea Pearl, *Causality*.

The ability to coax structure out of large datasets, particularly for difficult to program tasks, such as computer vision, speech recognition, and high-performance game playing, has led to significant successes of machine learning in a variety of real-world tasks, particularly using deep learning approaches. Broadly speaking, machine learning or data science is the process of constructing a probability distribution from samples, or equivalently being able to generate new samples from given samples that fool an expert discriminator (Goodfellow et al. 2014).

The fundamental difference between data science and imagination science is that the latter extends to realms far beyond the former: for example, imagination science addresses the problem of generating samples that are "novel", meaning they come from a distribution different from the one used in training. Imagination science also addresses the problem of causal reasoning to uncover simple explanations for complex events, and uses analogical reasoning to understand novel situations.

Can computers produce novel art like Basquiat's painting? Recent work on a variant of a generative adversarial network called CAN (for Creative Adversarial Network) (see Figure 2) shows that computers can be trained to produce images that are both art, as well as differ from standard styles, like impressionism or cubism. While CANs are a useful step forward, building on Berlyne's theory of novelty (Berlyne 1960), their architecture is currently specific to art, and not general enough to provide a computational framework for imagination. However, it does suggest one possible avenue to designing an Imagination Network architecture. Other extensions of GAN models, such as CycleGAN (Zhu et al. 2017), are suggestive, but such extensions are at present tailored to visual domains, and even in that circumscribed setting, only capable of specific generalizations (e.g., turning Monet styled watercolor paintings into what look like digital photographs of the original scene).

Most machine learning is based on the discovery and exploitation of statistical correlations from data, including approaches using parametric graphical model representations (Murphy 2013) or kernel-based non-parametric representations (Schölkopf and Smola 2002), and most recently, nonlinear neural net based models (Goodfellow, Bengio, and Courville 2016). Correlation, as has been pointed out many times, is not causation, however, and causal reasoning is one of the primary hallmarks of human imaginative reasoning (Pearl 2009). One of the primary rationales for causal reasoning is the need to provide comprehensible explanations, which will become increasingly important as autonomous systems play an ever larger role in society. A self-driving car that gets involved in an accident may be required to provide an explanation of its behavior, much as a human driver would, and such explanations often take on a causal form (see Figure 3).

A hallmark of imagination is the ability to reason about counterfactuals (Pearl 2009). The links between causal reasoning and imagination are explored from a probabilistic Bayesian perspective in (Walker and Gopnik 2013). Humans seek causal explanations because they want to understand the world in simple "cause-effect" relationships. They make analogies to interpret strange worlds, like the interior of an atom, in terms of worlds they understand, like the solar system, even though such analogies are imperfect. As Coase suggests above, humans desire interpretable explanations, even at the expense of fidelity to reality.

Imaginative Perception: Labels to Affordances

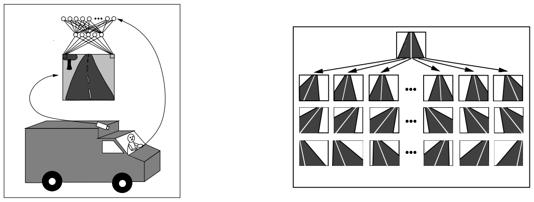
> "Although the field of A.I. is exploding with microdiscoveries, progress toward the robustness and flexibility of human cognition remains elusive. Not long ago, for example, while sitting with me in a cafe, my 3-year-old daughter spontaneously realized that she could climb out of her chair in a new way: backward, by sliding through the gap between the back and the seat of the chair. My daughter had never seen anyone else disembark in quite this way; she invented it on her own and without the benefit of trial and error, or the need for terabytes of labeled data." – Gary Marcus, *Artificial Intelligence is Stuck: Here's how to Move it forward.*, New York Times Sunday Review, July 29, 2017.

Alison Gopnik, a well known psychologist, in a recent article in Scientific American titled "Making AI More Human", marveled at the fact that "my five-year-old grandson, Augie, has learned about plants, animals and clocks, not to mention dinosaurs and spaceships. He also can figure out what other people want and how they think and feel. He can use that knowledge to classify what he sees and hears and make new predictions"(Gopnik 2017). One of the successes of machine learning, specifically deep neural networks (Goodfellow, Bengio, and Courville 2016), is object recognition. Performance on certain fixed datasets, such as Imagenet, has been steadily improving (Krizhevsky, Sutskever, and Hinton 2017). Under certain specific conditions, where large amounts of labeled datasets are available for narrowly defined tasks, deep learning approaches are able to exceed human level performance, a remarkable achievement. However, these results have to be interpreted with caution. There are as yet no well-defined procedures for extracting interpretable explanations from deep learning networks, and innocuous amounts of imperceptible noise appear sufficient to make a deep learning network guess the wrong label (Nguyen, Yosinski, and Clune 2015).

Children, such as Augie, generally do not excel at high-performance problem-solving in narrowly constrained problems, be it Imagenet or Go (Silver et al. 2016) or Jeopardy (Tesauro et al. 2014), but rather demonstrate extremely versatile *competence* at comprehending the world in all its multi-modal richness and dimensionality. AI researchers focusing almost entirely at superhuman performance on artificial datasets or puzzles are at risk of losing sight of what makes humans like Augie truly special. Challenge tasks in computer vision, speech recognition, and other areas, focus on the ability to label a particular object or scene (or transcribe a given dialog), where the emphasis is on *expert level ability given a statically defined task*. Children, in contrast, are capable of learning in a much more fluid manner, coping with significant variability between training and test distributions, and they seem to be able to learn quite effectively without requiring much explicit labeling.

Recent work is beginning to address the importance of imaginative causal reasoning in enabling neural net approaches to learn more effectively without labels (Stewart and Ermon 2017). However, a child interprets objects with an imaginative flexibility that lies far beyond what any AI system can accomplish today. To a child, a chair may serve as a hiding place, by crouching under it, or a stool, to retrieve another object placed beyond reach on a high table. In other words, using the terminology introduced by the psychologist James Gibson, imaginative thinking in perception revolves centrally around the core concept of "affordances": an object is perceived in terms of the actions it enables an agent to do, and not purely in terms of a descriptive label (Gibson 1977). A bed may suggest lying down to an adult, but to a child, it means many different things, including the ability to jump up and down on it, as children are apt to do. Affordances are also essential to the design of everyday appliances (Norman 2002).

What would it take to develop "Imagination Networks", an imaginative perceptual system that is able to interpret images with the same flexibility and richness of behavior that Gary Marcus' 3 year old child demonstrates, or the breadth of knowledge of Alison Gopnik's five-year-old grandson, Augie? For one, the ability to recognize and exploit affordances. Second, the ability to integrate perceptual information into an agent's goals, which are entirely a function of the agent's body. Affordances, like being able to get in and out of small openings in the back of a chair, depend on an agent's physical size and its capabilities. Simulated agents that function in video games, such as Atari, may have affordances that depend on their particular capabilities. Imaginative perception also plays a key role in other perceptual abilities, such as interpreting speech intonations and emotions, as well as body gestures. Affordances play a central role



(a) ALVINN learned to drive from human supervised data.

(b) Imagination in ALVINN

Figure 3: ALVINN (Pomerleau 1989) was an early attempt at building a self-driving vehicle that learned from observing human driving behavior. To accelerate learning, ALVINN employed a simple causal model of driving to imagine many hypothetical driving situations from each real experience.

in connecting objects with the actions they enable agents to undertake. The computation of affordances is an important objective for extending current work in deep learning for computer vision. Work on affordance recognition is still in its infancy, but important steps have been taken (Rome, Hertzberg, and Dorffner 2008). We discuss below how affordances can be learned by recognizing topological features in environments.

Imagination in Problem Creation

> "Perhaps no single phenomenon reflects the positive potential of human nature as much as intrinsic motivation, the inherent tendency to seek out novelty and challenges, to extend and exercise one's capacities, to explore, and to learn. Developmentalists acknowledge that from the time of birth, children, in their healthiest states, are active, inquisitive, curious, and playful, even in the absence of specific rewards" (Ryan and Deci 2000).

Much AI research has been focused on problem solving, but imagination provides the ability to do *problem creation*. A landmark early example was Doug Lenat's AM system (Lenat and Brown 1983), which was able to conjecture a large number of interesting theorems in elementary number theory (but lacked the ability to prove any of them). We now have many sophisticated ways to solve large Markov decision processes (Sutton and Barto 1998), but we lack the knowhow to create new MDPs. Deep reinforcement learning agents (Mnih et al. 2015) can play a precisely formulated Atari video game endlessly without getting bored, but the ability to create a new Atari video game remains completely out of reach of these systems. Yet, *game design* is a hallmark of imagination in human game developers, a skill whose success can be measured in billions of dollars of revenue.

It has long been recognized that reinforcement learning systems require the human to supply a reward function of the task. Yet, children seem capable of learning a wide variety of tasks, seemingly without explicit reward functions being supplied. One possible way humans acquire rewards is to learn them from observing the behavior of other humans, an approach referred to as *inverse reinforcement learning* (Ng and Russell 2000).

Another hallmark of imagination is the ability to get curious, to seek out novel situations, and to get bored solving the same problem repeatedly. Increasingly, many CEOs and managers have recognized is that excellence is often a by product of giving humans more autonomy in their work (Pink 2009). Intrinsic motivation has been studied in psychology for several decades (Ryan and Deci 2000), and now receiving increasing attention in machine learning (Singh, Barto, and Chentanez 2004).

Imagination in Language: Metaphors

> "I done wrestled with an alligator, I done tussled with a whale; handcuffed lightning, thrown thunder in jail; only last week, I murdered a rock, injured a stone, hospitalized a brick; I'm so mean I make medicine sick." – Muhammad Ali.

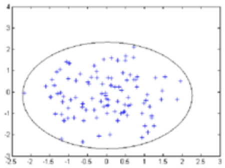
Metaphors play a crucial role in language (Lakoff and Johnson 1980), where phrases like "The stock market crashed" are apt to be used in everyday language. The creative use of metaphors in language, such as the above quotation from Muhammad Ali, shows the power of imagination in language. Recent successes in natural language processing, such as machine translation, build on deep learning sequence learning models, such as long short-term memory (LSTM) (Gers et al. 2002). However, the ability to undertake routine translation is far removed from the ability to creatively use language.

Recently, several techniques have emerged for mapping words into vectors, like `word2vec` (Mikolov, Yih, and Zweig 2013) or GLOVE (Pennington, Socher, and Manning 2014). Such word embedding systems can be trained on a large corpus of words, like Wikipedia, and produce continuous representations, which can be used to reason about linguistic relations (such as "Man is to Woman as King is to X"). A significant challenge remains in showing how word embedding techniques can be extended so that they can provide the basis for generating new metaphors or richly descriptive phrases of the sort quoted above, which lie at the heart of imaginative language use. What insights will enable AI to generate imaginative poetry?

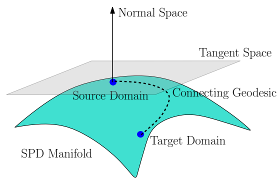
> "Beauty is truth, truth beauty, – that is all
> Ye know on earth, and all ye need to know." *Ode on a Grecian Urn*, John Keats.

Probabilistic Imaginative Models

Thus far, the paper has discussed the problem of designing imagination machines in a non-technical manner. In this last section, we briefly sketch out some ideas for how to develop a more rigorous mathematically-based theory of imagination. Machine learning is based on the core concept of a *probabilistic generative model* (PGM) (Murphy 2013), which concisely summarizes the distribution that generates both the training as well as the test datasets. Examples of PGMs include Gaussian mixture models, hidden Markov models, and Bayesian networks. Building on this concept,



(a) PGM



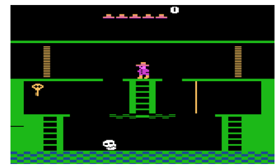
(b) PIM

Figure 4: This figure illustrates the difference between a probabilistic generative model PGM and a probabilistic imaginative model (PIM).

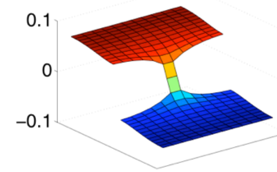
we introduce in this section the idea of *probabilistic imaginative models* (PIMs), which attempt to capture the essence of imagination science, and how it extends data science (see Figure 4). Broadly speaking, each PGM is represented by a point on a PIM, so the latter represents not a single dataset, but an entire universe of datasets, and its geometry is used to construct imagination routines.

In a probabilistic generative model, such as the simple Gaussian 2D ellipsoid illustrated in Figure 4a, samples from the training distribution are used to construct a generative model, in this case a 2D Gaussian, that with high likelihood produced the training data. In a probabilistic imaginative model, as shown in Figure 4b, each training set corresponds to a single point in a so-called *imagination space*. In this example, the space of covariance matrices is modeled as a *homogeneous space*, namely a curved Riemannian manifold that is acted upon by a continuous Lie group of invertible matrices. Every point on the manifold corresponds to a different dataset, and a different distribution. Given a training set of data points, and an unlabeled target set of data points, the process of constructing a PIM corresponds to finding a shortest path geodesic between the source and target datasets. It is possible to construct a new set of features from the training dataset that is shaped by the test samples, such as for example constructing the geometric mean (or sharp mean) of the source and target data covariances. Points along the geodesic correspond to *imagined datasets*, which have not previously been seen by the learner, but nonetheless represent valid possible points in the imagination space.

The homogeneous space constructed in Figure 4b is a special case of a much more general concept in mathematics called a *fiber bundle* (Husemller 1994), which represents a parameterized space that satisfies certain properties. For example, a Riemannian manifold is a fiber bundle, comprising of a *base space* of points, at each of which can be constructed a tangent space of *fibers*. Each such tangent space can be linked to other tangent spaces using the procedures developed in differential geometry. Interestingly, probabilistic imaginative models based on fiber bundles are significantly superior to linear vector-based approaches at capturing properties of linguistic relations (Mahadevan and Chandar 2015).



(a) Montezuma's Revenge



(b) Proto-value function

Figure 5: This figure illustrates how proto-value functions (PVFs) can be used to construct imagination spaces, and subsequently used to solve difficult control learning tasks.

Proto-Value Functions

We now turn to provide a second example of how to construct imagination spaces, based on the author's previous work on proto-value functions (PVFs) (Mahadevan 2005). Proto-value functions are task-independent value functions that are constructed from a reinforcement learning agent's random trial and error exploration through a state (action) space. Unlike pre-defined bases, like the radial basis function (RBF) or CMAC, PVFs adapt to the nonlinear geometry of a state (action) space, as shown in Figure 5b. In this example, the particular PVF represents an eigenfunction of the graph Laplacian operator defined on the space of all functions on the (discrete) state space of a environment with two rooms connected by a door. The PVF shows clearly the geometry of the space, and the bottleneck that exists between the two rooms. Figure 5a shows one of the Atari video games called Montezuma's Revenge, which the standard DQN deep RL approach completely failed to solve (Mnih et al. 2015). However, recent work (Machado, Bellemare, and Bowling 2017) has shown that PVFs can be used to construct *eigenpurposes*, intrinsically rewarded behavior where each PVF is treated as an internally generated task-independent value function, using which a deep Q-learner can bootstrap itself to solve this difficult Atari video game. Mathematically, the space of all possible eigenfunctions on a state (action) space can be decomposed into a *flag manifold* (Monk 1959), a special type of homogeneous space that is also a fiber bundle. A flag manifold is a nested series of subspaces, where each subspace is defined as the span of a corresponding set of PVFs.

Acknowledgments

This paper was written while the author was on leave from U.Mass, Amherst, serving as the Director of the SRI Artificial Intelligence Center in Menlo Park, CA. The author is indebted to SRI for their enthusiastic support of this research. The author also thanks several members of the AI community who provided helpful feedback on earlier drafts, including Nils Nilsson, Judea Pearl, and Stuart Russell.

References

Alexander, I. 2001. *How to build a mind: Towards machines with imagination*. Columbia University Press.

Asma, S. 2017. *The Evolution of Imagination*. University of Chicago Press.

Berlyne, D. 1960. *Conflict, Arousal and Curiosity*. McGraw-Hill.

Elgamman, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. CAN: Creative adversarial networks generating “art” by learning about styles and deviating from style “norms”. *Arxiv*.

Gers, F.; Pérez-Ortiz, J.; Eck, D.; and Schmidhuber, J. 2002. Learning context sensitive languages with LSTM trained with Kalman filters. In Dorronsoro, J., ed., *Artificial Neural Networks – ICANN 2002 (Proceedings)*, 655–660. Berlin: Springer.

Gibson, J. J. 1977. The theory of affordances. In Shaw, R., and Bransford, J., eds., *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. Erlbaum. 67–82+.

Goodfellow, I.; Bengio, Y.; and Courville, A. C. 2016. *Deep Learning*. Adaptive computation and machine learning. MIT Press.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2672–2680.

Gopnik, A. 2017. Making AI more human. *Scientific American* 62–65.

Husemller, D. 1994. *Fiber Bundles*. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60(6):84–90.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lenat, D. B., and Brown, J. S. 1983. Why AM and Eurisko appear to work. In *AAAI*, 236–240.

Machado, M. C.; Bellemare, M. G.; and Bowling, M. H. 2017. A Laplacian framework for option discovery in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2295–2304.

Mahadevan, S., and Chandar, S. 2015. Reasoning about linguistic regularities in word embeddings using matrix manifolds. *CoRR* abs/1507.07636.

Mahadevan, S. 2005. Proto-value functions: Developmental reinforcement learning. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML ’05, 553–560. New York, NY, USA: ACM.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, 746–751.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Monk, D. 1959. The geometry of flag manifolds. *Proceedings of the London Mathematical Society*.

Murphy, K. P. 2013. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press.

Ng, A. Y., and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML ’00, 663–670. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ng, A. 2016. Hiring your first chief AI officer. *Harvard Business Review* 2:100–107.

Nguyen, A. M.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 427–436.

Norman, D. A. 2002. *The Design of Everyday Things*. New York, NY, USA: Basic Books, Inc.

Pascanu, R.; Li, Y.; Vinyals, O.; Heess, N.; Buesing, L.; Racanire, S.; Reichert, D.; Weber, T.; Wierstra, D.; and Battaglia, P. 2017. Learning model-based planning from scratch. *Arxiv*.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, 1532–1543.

Pink, D. H. 2009. *Drive: the surprising truth about what motivates us*. New York: Riverhead Books.

Pomerleau, D. A. 1989. Alvinn: An autonomous land vehicle in a neural network. In Touretzky, D. S., ed., *Advances in Neural Information Processing Systems 1*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 305–313.

Rome, E.; Hertzberg, J.; and Dorffner, G., eds. 2008. *Towards Affordance-Based Robot Control - International Seminar, Dagstuhl Castle, Germany, June 5-9, 2006. Revised Papers*, volume 4760 of *Lecture Notes in Computer Science*. Springer.

Ryan, R. M., and Deci, E. L. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55(1):68–78.

Schölkopf, B., and Smola, A. J. 2002. *Learning with Kernels*. MIT Press.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Singh, S. P.; Barto, A. G.; and Chentanez, N. 2004. Intrinsically motivated reinforcement learning. In *NIPS*, 1281–1288.

Stewart, R., and Ermon, S. 2017. Label-free supervision of neural networks with physics and domain knowledge. In *Proceedings of the AAAI Conference*.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Taylor, M. 2013. Transcending time, place, and/or circumstance: An introduction. In Taylor, M., ed., *Oxford Handbook of the Development of Imagination*. Oxford University Press. 3–10.

Tesauro, G.; Gondek, D.; Lenchner, J.; Fan, J.; and Prager, J. M. 2014. Analysis of Watson’s strategies for playing jeopardy! *CoRR* abs/1402.0571.

Walker, C., and Gopnik, A. 2013. Causality and imagination. In Taylor, M., ed., *Oxford Handbook of the Development of Imagination*. Oxford University Press. 342–358.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*.