Chapter 1

Artificial Intelligence in the Humanities: Wolf in Disguise, or Digital Revolution?

Arend Hintze
Dalarna University

Jorden Schossau
Michigan State University

Introduction

Artificial Intelligence, with its ability to machine learn coupled to an almost human-like understanding, sounds like the ideal tool to the humanities. Instead of using primitive quantitative methods to count words or catalogue books, current advancements promise to reveal insights that otherwise could only be obtained by years of dedicated scholarship. But are these technologies imbued with intuition or understanding, and do they learn like humans? Are they capable of developing their own perspective, and can they aid in qualitative research?

In the 80s and 90s, as home computers were becoming more common, Hollywood was sensationalizing the idea of smart or human-like Artificial Intelligent machines (AI) through movies such as *Terminator*, *Blade Runner*, *Short Circuit*, and *Bicentennial Man*. At the same time, the home experience of personal computing highlighted the difference between Hollywood intelligent machines and the reality of how "dumb" machines really were. Home, or even industry machines, could not answer simple natural language questions of anything but the simplest of complexity. Instead, users or programmers needed to painstakingly implement an algorithm to address their question. Then, the user was required to wait for the machine to slavishly follow each instruction that was programmed while hoping that whoever entered the instructions did

not make a mistake. Despite the Hollywood intelligent machines sensation, people understood that computers did not and could not think like humans, but that they do excel at performing repetitive tasks with extreme speed and fidelity. This shaped the expectations for interacting with computers. Computers became efficient tools that required specific instruction in order to achieve a desired outcome.

Computational technology and user experience drastically changed over the next 20 years. Technology became much more intuitive to use while it also became much more powerful at handling large data sets. For instance, Google can return search results for websites as a response to even the silliest or sparsest request, with a decent chance that the results are relevant to the question asked. Did you read a manual before you used your smartphone, or did you like everyone else just "figure it out"? Or, as a consequence of modern-day media and its on-demand services, children ask to skip a song playing through radio broadcast. The older technologies quickly feel archaic.

These technological advancements go hand in hand with the developments in the field of machine learning and artificial intelligence. The automotive industry is on the cusp of fully self-driving cars. Electronic assistants are not only keeping track of our dates and responding to spoken language, they will also soon start making our appointments by speaking to other humans on our behalf. Databases are getting new voice-controlled intuitive interfaces, changing a typical incomprehensible "`SELECT AVG(salary) FROM employeeList WHERE yearHired > 2012;`" to a spoken "Average salary of our employees hired after 2012?"

Another phenomenon is the trend in many disciplines to go from "qualitative" to "quantitative" research, or to think about the "system" rather than the "components." The field that probably experienced this trend first was biology. While obviously descriptive about species of organisms, biologists also always wanted to understand the mechanisms that drive life on earth spanning micro to macro scales. Consequently, a lot is known about the individual chemical components that constitute our metabolism, the components that drive cell division and DNA replication, and which genes are involved in, for example, developmental processes. However, in many cases, our scientific knowledge only covers single functions of single components. In the context of the cell, the state of the organism and how other components interact matters a lot. Cancer, for example, cannot be explained by a single mutation on a single gene but involves many complex interactions (Hanahan and Weinberg 2011). Ecosystems don't collapse because a single insect dies, but because indirect changes in the food chain interact in complex ways (for a review of the different theories, see Tilman 1996). As a result, systems biology emerged. Systems biologists use large data sets and are often dependent on computer models to understand phenomena on the systems level.

The field of Bioinformatics is one such example of an entire field that emerged as a result of using computers to study entire systems that were otherwise humanly intractable. The human genome project to sequence the complete human genome finished in 2003, a time when our consumer data storage was limited by the amount of data that fit on a DVD (4.9 GB). While the human genome fits on a DVD, the data that came from the sequencing machines was much larger. Short repetitive sequences first needed assembly, which at that time was a high-performance computing task.

Other fields have since undergone their own computational revolutions, and now the humanities begin their computational revolution. Computers have been a part of core library infrastructure and experience for some time now, by cataloging entries in a database and allowing intuitive user exploration of that database. However, the digital humanities go beyond this (Fitz-

patrick 2012). The ability to analyze (crawl) extremely large corpora of different sources, monitor the internet using the Internet of Things as large sensor arrays, and detect patterns by using sophisticated algorithms can each produce a treasure trove of quantitative data. Until this point these tasks could only be described or analyzed qualitatively.

Additionally, artificial intelligence promises models of the human mind (Yampolskiy and Fox 2012). Machine learning allows us to learn from these data sets in ways that exceed human capabilities, while an artificial brain will eventually allow us to objectively describe a subjective experience (through quantifying neural activations or positively and negatively associated memories). This would ultimately close the gap between quantitative and qualitative approaches by allowing an inspection of experience.

However, this bridging between quantitative and qualitative methods causes a possible tension for the humanities, which historically defines itself by qualitative methodologies. When qualitative experiences or responses can be finely quantified, such as sadness caused by reading a particular passage, or the curiosity caused by viewing certain works of art, then the field will undergo a revolution. When this happens, we will be able to quantify and discuss how sadness was learned by reading, or how much surprise was generated by viewing an artwork.

This is exactly the point where the metaphors break down. Current computational models of the mind are not sophisticated enough to allow these kinds of inferences. Machine learning algorithms work well for what they do but have nothing to do with what a person would call learning. Artificial intelligence is a broad encompassing field. It includes methods that might have appeared to be magic only a couple of years ago (such as generative adversarial networks). Algorithmic finesse resulting from these advances is capable of beating humans in chess (Campbell, Hoane Jr, and Hsu 2002), but it is only a very specialized algorithm that has nothing to do with the way humans play or learn chess. This means we are back to the problem we had in the 80s. Instead of being disappointed by the difference between modern technology and Hollywood technology, we are disappointed by the difference between modern technology and the experience implied by the labels given to those technologies. Applying misnomer terminology, such as "smart," "intelligent," "search," and "learning" to modern technologies that have little to do with those terms is misleading. It is possible that such technology was deliberately branded with these terms for the improved marketing and sales, effectively redefining them and obscuring their original meaning. Consequently, we again are disappointed by the mismatch of the expectations of our computing infrastructure and the reality of our experiences.

The following paragraphs will explore current Machine Learning and Artificial Intelligence technologies, explain how quantitative or qualitative they really are, and explore what the possible implications for future Digital Humanities could be.

Learning: Phenomenon versus Mechanism

Learning is an electrochemical process that involves cells, their genetic makeup, and how they are interconnected. Some interplay between external stimuli and receptor proteins in specialized sensor neurons leads to electrochemical signals propagating over a network of interconnected cells, which themselves respond with physical and genetic changes to said stimuli, probably also dependent on previous stimuli (Kandel, Schwartz, Jessel 2000). This concoction of elaborate terms might suggest that we know in principle which parts are involved and where they are, but we are far from an understanding of the learning mechanism. The description above is as generic as saying that a city functions because cars drive on streets. Even though we might know a lot

about long-term potentiation or the mechanism of neurons which fire together wiring together (aka Hebbian learning), neither of these processes actually mechanistically explains how learning works. Neuroscience, neurophysiology, and cognitive science have not been able to discover this complete process in such a way that we can replicate it, though some inroads are being made (El-Boustani et al. 2018). Similarly, we find promising new interdisciplinary efforts like "Cognitive computational neuroscience" that try to bridge the gap between neuro- and cognitive science and computation (Kriegeskorte and Douglas 2018). So, unfortunately, while the components involved can be identified, the question about "how learning works" cannot be answered mechanistically.

However, a lot is known about the phenomenon of learning. It happens during the lifetime of an organism. What happens between the lifetimes of related organisms is an adaptive process called evolution: inheritance, variation, and natural selection over many generations up to 3.5 billion years here on Earth enabled populations of organisms to succeed in their environments in any way they could. Evolutionary forces found ways for organisms to adapt to their environment during their own lifetimes. While this can take many forms, such as storing energy, seeking shelter, having a fight or flight response, it has led to the phenomenon we now call learning. Instead of discussing the diversity of learning in the animal kingdom, we will discuss the richest example: human learning.

Here, learning is defined as the cognitive adaptation to external stimulus. The phenomenon of learning can be observed as an increase in performance over time. Learning makes the organism better at doing something. In humans, because we have language and a much higher degree of abstract thinking, an improvement in performance can be facilitated very quickly. While it takes time to learn how to juggle, the ability to find the mean of a series of samples can be quickly communicated by reading Wikipedia. Both types of lifetime adaptations are called learning. However, these lifetime adaptations are facilitated by two different cognitive processes: explicit or implicit learning. [1] Explicit learning—or episodic memory—is fact-based memory. What you did yesterday, what happened in your childhood, or the list of things you should buy when you go shopping, are all memories. Currently, the engram theory best explains this mechanism (Poo et al. 2016 elaborates on the origins of the term). Explicit memory can be retrieved relatively easily and then used to inform future decisions: "Press the green button if the capital of Italy is Paris, otherwise press the red." The rate of learning for explicit memory can be much higher than for implicit memory, and it can also be communicated more quickly. Abstract communication, such as "I saw a wolf" allows us to transfer the experience of seeing a wolf quickly to other individuals, even though their evoked explicit memory might not be identical to ours.

Learning by using implicit memory—sometimes called procedural memory—is facilitated by much slower processes (Schacter, Chiu, and Ochsner 1993). It is generally based on the idea that learning is a combination of expectation, observation or action, and internal model changes. For example, a recovering hospital patient who has suffered a stroke is handed an apple. In this exchange, the patient forms an expectation of where his hand will be to accept the apple. He engages his muscles to move his forearm and hand to accept the apple, which is his action. Then the patient observes that his arm did not arrive at the correct position (due to neurological damage). This discrepancy between expectation and action-outcome drives internal changes so that the patient's brain learns how to adequately control their arm. Presumably, everything considered a skill is based on this process. While very flexible, this form of memory is not easily communicated nor fast to acquire. For instance, while juggling can be described it cannot be communicated in

[1] There are more than these two mechanisms, but these are the two major ones.

such a way that it enables the recipient to juggle without additional training.

This description of explicit and implicit learning is an amalgamation of many different hypotheses and observations. Also, these processes are not as well segregated in practice as outlined here. What is important is what these two learning mechanisms are based on: observations lead to memory, and internal predictions together with exploration lead to improved models about the world. Lastly, these learning processes only exist in organisms because they previously conferred an evolutionary advantage: Organisms that could memorize and then act on those memories had more offspring than those that did not. This interaction of learning and evolution is called the Baldwin effect (Weber and Depew 2003). Organisms that could explore the environment, make predictions about it, and use observations to optimize their internal models were similarly more capable than organisms that could not.

Machines do not Learn; They are Trained

Now prepared with a proper intuition about learning, we can turn our attention to machine learning. After all, our intuitions should be meaningful in the computational domain as well, because learning always follows the same pattern. One might be disappointed when looking over the table of contents of a machine learning book and find only methods for creating static transformation functions (see Russell and Norvig 2016, one of the putative foundations of machine learning and AI). There will typically be a distinction between supervised and unsupervised learning, between categorical and continuous data, and maybe a section about other "smart" algorithms. You will not find a discussion about implicit and explicit memory, let alone methods for implementing these concepts. So, if these important sections in our imaginary machine learning book do not discuss the mechanisms of learning, then what are they discussing?

Unsupervised learning describes algorithms that report information based on associations within the data. Clustering algorithms are a popular example of unsupervised learning. These use similarity between data points to form and report on distinct groups of data. Clustering is a very important method but is only a well-designed algorithm that is not adaptive.

Supervised learning describes algorithms that refine a transformation function to convert from a certain input to a certain output. The idea is to balance specific and general refining such that the transformation function correctly transforms all known examples but generalizes enough to work well on new variations. For example, we would like the machine to transform image data into textual labels, such as "house" or "car." The input is an image and the output is a label. The input image data are provided to the machine, and small adjustments to the machine's function are made depending on how well it provided the correct output. Many iterations later ideally will result in a machine that can transform all image data to correct labels, and even operate correctly on new variations of images not provided before. Supervised learning is extremely powerful and is yet to be fully explored. However, supervised learning is quite dissimilar to actual learning.

A common argument is that supervised learning uses feedback in a "student-teacher" paradigm of making changes with feedback until proper behavior is achieved, so it could be considered learning. But this feedback is external, objective, and not at all similar to our prediction and comparison model that, for instance, operates without an all-knowing oracle whispering "good" or "bad" into our ears. Humans and other organisms instead compare predictions with outcomes, and the choices are driven by an intersection of desire and prediction.

What seems astonishing is the diverse and specialized capabilities that these two rather simple types of computation, clustering and classification, can produce. Their economic impact is enor-

mous, and we are still finding new ways to combine neural networks and exploit deep learning techniques to create amazing data transformations, such as deep fake videos. But so far, each astounding example of AI, through machine learning or some other method, is not showcasing all these capabilities as one machine, but instead each as an independently achieved computational marvel. Each of these examples does only exactly what it was trained to do in a narrow domain and no more. Siri, or any other voice assistant for that matter, does not drive a car (López, Quesada, and Guerrero 2017), Watson does not play chess (Ferrucci et al. 2013), and Google Alpha Go cannot understand spoken language (Gibney 2016). Even hybrid approaches, such as combining speech recognition, chess playing, and autonomous driving, would only be a combination of specialty strategies, not a trained entity from the ground up.

Modern machine learning gives us an amazing collection of very applicable, but extremely specialized, computational tools that may be customized to particular data sets, but the resulting machines do not learn autonomously as you or I do. There are cutting edge technologies, such as so-called neuromorphic chips (Nawrocki, Voyles, and Shaheen 2016) and other computational brain models that more closely mimic brain function, but they are not what has been sensationalized in the media as machine learning or AI, and they have yet to showcase competence on difficult problems competitive with standard supervised learning.

Curiously, many people in the machine learning community defend the term "learning," arguing there is no difference between learning and training. In traditional machine learning, the trained algorithm is deployed as a service after which it no longer improves. If the data set ever changes, then a new training set including correct labels needs to be generated and a new training phase initiated. However, if the teacher can be forever bundled with the learner and training continued during the deployment phase, even on new never-before-seen data, then indeed the delineation between learning and training is far less clear. Approaches to such lifelong learning exist, but they struggle with what is called *catastrophic forgetting*—the phenomenon that only the most recent experiences are learned at the expense of older ones (French 1999). This is the objective for Continuous Delivery for machine learning. Unfortunately, creating a new training set is typically the most expensive endeavor for standard supervised machine learning development. Adequate training then becomes difficult or impossible without involving thousands or millions of human inputs to keep up with training and using the online machine on an ever-evolving data set. Some have tried to use such "human-in-the-loop" methods, but the resulting machine then becomes only a slight extension of the humans who are forever caught in the loop. Is it an intelligent machine, or a human trapped in a machine?

To combat this problem of generating the training set, researchers altered the standard supervised learning paradigm of flexible learner and rigid teacher to make the teacher likewise flexible to generate new data, continually probing the bounds of the student machine. This is the method of Generative Adversarial Networks, or GANs (Goodfellow et al. 2014). The teacher generates training examples and the student discerns between those generated examples and the original labeled training data. After many iterations, the teacher is improved to better fool the student, and the student is improved to better discern generated training data. As amazing as they are, GANs only partially mitigate the problematic requirement for human-labeled training data, because GANs can only mimic a known labeled distribution. If that distribution ever changes, then new labeled data must be generated, and again we have the same problem as before. Unfortunately, GANs have been sensationalized as magic, and public and hobbyist expectation is that GANs are a way toward much better artificial intelligence. Disappointment is inevitable because GANs only allow us to explore what it would be like to have more training data from the same

data sets we were using before.

These expectations are important for machine learning and AI. We are very familiar with learning, to the point where our whole identity as human could be generously defined as the result of being a monkey with an exceptional proclivity for learning. If we now approach AI and machine learning with expectations that these technologies learn as we do, or are an equally general-purpose intelligence, then we will be bitterly disappointed. The best example of such discrepancy is how easily neural networks trained by deep learning can be fooled. Images that are seemingly identical and differ only by a few pixels are grossly misclassified, a mistake no human would make (Nguyen, Yosinski, and Clune 2015). Fortunately, we know about these biases and the possible shortcomings of these methods. As long as we have the right expectations, we can take their flaws into account and still enjoy the prospects they provide.

Trained Machines: Tool or Provocation?

On one side we have the natural sciences characterized by hypothesis-driven experimentation reducing reality to an abstract model of causal interactions. This approach can inform us about the consequences of our possible actions, but only as far in the future as the model can adequately predict. With machine learning and AI, we can move this temporal horizon of prediction farther into the future. While weather models might still struggle to predict precipitation 7 days in advance, global climate models predict in detail the effects of global warming in 100 years. But these models are nihilist, void of values, and cannot themselves answer the question if humans would prefer to live in one possible future or another. Is sunshine better than rain? The humanities, on the other hand, are home to exactly these problems. What are our values? How do we understand what is essential? Now that we know the facts, how should we choose? Do we speak for everyone? The questions seem to be endless, but they are what makes our human experience so special, and what separates the humanities from the sciences.

Labels—such as learning or intelligence—are too easily anthropomorphized. A technology branded in this way suggests human-like properties: intelligence, common sense, or even subjective opinion. From a name like "deep learning" we expect a system that develops a deep and intuitive understanding with insights more profound than our own. However, these systems do not provide an alternative perspective, but as explained above, are only as good or as biased as the scientist selecting their training data. Just because humans and machine learning are both black boxes in the sense that their inner workings are opaque, does not mean they share other qualities. For instance, having labeled the ML training process as "learning" does not imply that ML algorithms are curious and learn from observations. While these new computerized quantitative measures might be welcomed by some scholars, there will be others who view it as an existential threat to the very nature of the humanities. Are these quantitative methods sneaking into the humanities disguised by anthropomorphic terms like a wolf shrouded in a sheep's fleece? From this viewpoint, having the wrong expectations is not only provoking a disappointment, but flooding the humanities with sophisticated technologies that dilute and muddy the nature of qualitative research that makes the humanities special.

However, this imminent clash between quantitative and qualitative research also provides a unique opportunity. Suppose there is a question that can only be answered subjectively and qualitatively. If so, it would define a hard boundary against the aforementioned reductionism of the purely causal quantitative approach. At the same time, such a boundary presents the perfect target for an artificially intelligent system to prove its utility. If a computational human analog

can be created, then it must be capable of performing the same tasks as a humanities researcher. In other words, it must be able to answer subjective and qualitative questions, regardless of its computational and quantitative construction. Failing at such a task would be equivalent to failing the famous Turing test, thereby proving the AI is not yet human-like enough. In this way, the qualitative nature of the humanities poses a challenge—and maybe a threat—to artificially intelligent systems. While some might say the threat is mutual, past successes of interdisciplinary research suggest otherwise: The digital humanities could become the forefront of AI research.

Beyond machine training, towards general purpose intelligence

Currently, machines do not learn but must be trained, typically with human-labeled data. ML algorithms are not smart as we are, but they can solve specific tasks in sophisticated ways. Perhaps sentience will only be a product of enough time and training data, but the path to sentience probably requires more than time and data. The process that gave rise to human intelligence was evolution. This opportunistic process optimized brains over endless generations to perform ever-changing tasks, and it is the only known example of a process that resulted in such complex intelligence. None of the earlier described computational methods even remotely follow this paradigm: Researchers designed ad hoc algorithms that solved well-defined problems. The next iteration of these methods is either an incremental improvement of existing code, a new methodological invention, or an application to a new data set. These improvements do not compound to make AI tools better generalists, but instead contribute to the diversity of the existing tools.

One approach that does not suffer from these shortcomings is neuro-evolution (Floreano, Dürr, and Mattiussi 2008). Currently, the field of Neuroevolution is in its infancy, but finding new and creative solutions to otherwise unsolved problems, such as controlling robots driving cars, is a popular area of focus (Lehman et al. 2020). At the same time, memory formation (Marstaller, Hintze, and Adami 2013), information integration in the brain (Tononi 2004), and how systems evolve the ability to learn (Sheneman, Schossau, and Hintze 2019) are also being researched, as they are building blocks of general purpose intelligence. While it is not clear how thinking machines will ultimately emerge, they are on the horizon. The dualism of a quantitative system that can be subjective and understand the qualitative nature of existence makes it a strange artifact that cannot be ignored.

References

Campbell, Murray, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. "Deep Blue." *Artificial Intelligence* 134 (1–2): 57–83.

El-Boustani, Sami, Jacque P K Ip, Vincent Breton-Provencher, Graham W Knott, Hiroyuki Okuno, Haruhiko Bito, and Mriganka Sur. 2018. "Locally Coordinated Synaptic Plasticity of Visual Cortex Neurons in Vivo." *Science* 360 (6395): 1349–54.

Ferrucci, David, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. 2013. "Watson: Beyond Jeopardy!" *Artificial Intelligence* 199: 93–105.

Fitzpatrick, Kathleen. 2012. "The Humanities, Done Digitally." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 12–15. Minneapolis: University of Minnesota Press.

Floreano, Dario, Peter Dürr, and Claudio Mattiussi. 2008. "Neuroevolution: From Architectures to Learning." *Evolutionary Intelligence* 1 (1): 47–62.

French, Robert M. 1999. "Catastrophic Forgetting in Connectionist Networks." *Trends in Cognitive Sciences* 3 (4): 128–35.

Gibney, Elizabeth. 2016. "Google AI Algorithm Masters Ancient Game of Go." *Nature News* 529 (7587): 445.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, edited by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, 2672–80. N.p.: Neural Information Processing Systems Foundation.

Hanahan, Douglas, and Robert A Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5): 646–74.

Kandel, Eric R, James H Schwartz, and Thomas M Jessell. 2000. *Principles of Neural Science*. 4th ed. New York: McGraw-Hill.

Kriegeskorte, Nikolaus, and Pamela K Douglas. 2018. "Cognitive Computational Neuroscience." *Nature Neuroscience* 21: 1148–60.

Lehman, Joel et al. 2020. "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities." *Artificial Life* 26 (2): 274-306.

López, Gustavo, Luis Quesada, and Luis A Guerrero. 2017. "Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces." In *International Conference on Applied Human Factors and Ergonomics*, edited by Isabel L. Nunes, 241–50. Cham: Springer.

Marstaller, Lars, Arend Hintze, and Christoph Adami. 2013. "The Evolution of Representation in Simple Cognitive Networks." *Neural Computation* 25 (8): 2079–2107.

Nawrocki, Robert A, Richard M Voyles, and Sean E Shaheen. 2016. "A Mini Review of Neuromorphic Architectures and Implementations." *IEEE Transactions on Electron Devices* 63 (10): 3819–29.

Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2015. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 427–36. N.p.: IEEE.

Poo, Mu-ming et al. 2016. "What Is Memory? The Present State of the Engram." *BMC Biology* 14: 1-18.

Russell, Stuart J, and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Malaysia: Pearson Education Limited.

Schacter, Daniel L, C-Y Peter Chiu, and Kevin N Ochsner. 1993. "Implicit Memory: A Selective Review." *Annual Review of Neuroscience* 16 (1): 159–82.

Sheneman, Leigh, Jory Schossau, and Arend Hintze. 2019. "The Evolution of Neuroplasticity and the Effect on Integrated Information." *Entropy* 21 (5): 1-15.

Tilman, David. 1996. "Biodiversity: Population versus Ecosystem Stability." *Ecology* 77 (2): 350–63.

Tononi, Giulio. 2004. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5: 1–22.

Weber, Bruce H, and David J Depew. 2003. *Evolution and Learning: The Baldwin Effect Reconsidered*. Cambridge, MA: Mit Press.

Yampolskiy, Roman V, and Joshua Fox. 2012. "Artificial General Intelligence and the Human Mental Model." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Ammon H. Eden, James H. Moor, Johnny H. Søraker, and Erik Steinhart, 129–45. Heidelberg: Springer.