2020

Robot Bites Swan

Sergio Da Silva, *Federal University of Santa Catarina*

Open Access Library Journal
2020, Volume 7, e6117
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Robot Bites Swan

Sergio Da Silva

Department of Economics, Federal University of Santa Catarina, Florianopolis, Brazil
Email: professorsergiodasilva@gmail.com

How to cite this paper: Da Silva, S. (2020) Robot Bites Swan. *Open Access Library Journal*, **7**: e6117.
https://doi.org/10.4236/oalib.1106117

Received: January 29, 2020
Accepted: February 16, 2020
Published: February 19, 2020

Abstract

I evaluate the claim that black swans can be predicted by Bayesian machine learners and notice that it is "gray" that is being taken into account instead. Additionally, I offer a reminder that the socioeconomic branch of statistical physics also strives to spot gray swans.

Subject Areas

Psychology

Keywords

Black Swan, Problem of Induction, Artificial Intelligence, Machine Learning

1. Swan Bites Robot

In *The Black Swan*, Nassim Taleb forcefully made the point that if you only ever see white swans, you think the probability of ever seeing a black swan is zero [1]. This addresses the problem of induction, which is the logical-philosophical extension of the black swan problem [1]. Induction refers to methods that infer or predict that occurrences of which we have had no experience resemble those of which we have had experience [2]. However, just by arguing that an occurrence has always or usually been reliable in the past cannot be proved deductively. Because it is logically possible for the prediction to be false while past evidence is true, the evidence does not conclusively establish the truth of the prediction [3].

David Hume demonstrated that the problem of induction is insoluble. Karl Popper then held that induction is superfluous and has no place in the logic of science. Science is just a deductive process where hypotheses are tested through deriving particular observable consequences. No hypothesis can ever be confirmed. It can only be falsified and then rejected or tentatively accepted in the absence of falsifiability [4]. Popper's view on induction is one of the most influential and Taleb draws on it.

Artificial intelligence (AI) has reopened the problem of induction. Under the more recent influence of Taleb, the most common objection to machine learning—the bulk of AI—is to invoke the swan example: "No matter how smart your algorithm, there are some things it just can't learn" [5]. Computer Scientist Pedro Domingos remarks that while some events cannot be predictable, others can. The first duty of a machine learner is to disentangle them, while its ultimate goal is to learn everything that can be known, which is a wider domain than Taleb and the critics of AI imagine [5].

2. Robot Bites Swan

Domingos dares to hold that learning algorithms can predict never-before-seen events, and that is exactly what machine learning is all about. The probability of a never-before-seen black swan comes from the proportion of known species that turned out to have black varieties [5]. This stance goes "meta", and going meta is just "Simonyi's law" at work—"Anything that can be done could be done meta" [6].

Consider an example from economics of going meta in theorizing—the taxonomy of market efficiency in "weak form", "semi-strong form" and "strong form". Weak-form market efficiency refers to the past series of prices only, by positing that this cannot be used to predict future prices. Semi-strong-form market efficiency restates it by going meta—by adding a further layer of information to that of the time series of prices: All new public information now counts. And the strong form goes meta one more degree above: All information, whether public or private, counts.

The time series of observations of white swans still cannot predict a never-before-seen black swan. I call this the "weak-form problem of induction". And undoubtedly it remains intact, as a black swan cannot be predicted from a time series of only white swans. Domingos goes meta by adding information not only from swans, but also from all public information regarding, say, other white birds that turned out black. In this semi-strong-form problem of induction, black swans can be predicted due to an extension of the original information set. But more data are not enough. Here an extra rule is needed, and Bayes' theorem can fill the gap, as I show below. (Actually, if a black swan turned out to be predictable, it is because it was "gray" from the start; more on this next.)

This is machine learning's practical approach to the problem of induction, and it seems to be working, as exemplified later. We can even imagine a strong-form problem of induction, where robots can learn how to track private (or currently unknown) information on all white birds, past and present, in addition to the public (currently known) information.

Taleb acknowledges in Part III of *The Black Swan* [1] the existence of "gray swans", and this consent makes my interpretation of Domingos' stance—that many supposedly black swans are in fact gray—credible, because what can be known is a broader realm than Taleb envisions.

3. Silicon Beats Carbon

A related objection to AI it that "data can't replace human intuition" [5]. Domingos argues that it is the other way around: Human intuition cannot replace data. Indeed, human intuition is cognitively bounded [7]. In Chapter 21 of *Thinking, Fast and Slow* [7], Daniel Kahneman shows plenty of examples where a machine beats human intuition and also discusses the psychological hostility to algorithms. He concludes that whenever we can replace human judgment by a formula, we should at least consider it because many judgments thought of as complex and subtle cannot outperform a single combination of scores.

Hostility can turn to blindness. You may not know it, but machine learning is all around us. In *The Master Algorithm*'s preface [5], Domingos offers many astonishing examples. Watson the computer was a *Jeopardy*! champion, Deep Blue beat human chess grandmaster Garry Kasparov, and AlphaGo—a reinforcement learner with neural networks from DeepMind—beat a human player at Pong and other arcade games. We all witnessed the progress made by Google Translate in recent years and became heavy users of it. "Machine learning is remaking our world" [5]. The practical approach to the induction problem seems to be working by solving daunting problems.

4. Gray Swan in the Church of Reverend Bayes

Domingos asserts that Bayesian networks can compute the probabilities of extremely unusual states, including states that were never observed before [5]. The Bayesian master algorithm is used by one of the five tribes of machine learning (more on this later). So the claim is that machine learning can predict a black swan. As I hinted before, machine learning actually spots a gray swan. The elusive black swan escapes by definition because the problem of induction in weak form remains intact.

Bayesian networkers interpret Bayes' theorem

$$P(A|B) = P(A)P(B|A)/P(B)$$

as

$$P(\text{cause}|\text{effect}) = P(\text{cause})P(\text{effect}|\text{cause})/P(\text{effect}).$$

If we observe an effect that would happen even without the cause, this is not evidence of the cause being present. Bayes' theorem considers this, because $P(\text{cause}|\text{effect})$ is not the same as the "prior" probability of the effect $P(\text{effect})$ (the prior is the probability in the absence of any knowledge of the causes). But Bayes' theorem goes further by finding that the more likely a cause is a priori, the more likely it should be a posteriori [5]. Observe that this adds another layer of information to the problem of induction.

Bayesian networks are highly competent expectation generators that do not have to comprehend what they are doing [6]. This seems unbelievable if you think consciousness comes first and competence second. Thinking so, you are unaware of what Daniel Dennett [6] calls Darwin's strange inversion of reason-

ing—competence occurs in nature even in the absence of consciousness, and consciousness evolves from competence (and this justifies the fear of a robot takeover, which is an issue I discuss later). Domingos instantiates Darwin's inversion—competence without comprehension—by observing it has been occurring in machine learning [6]. This should not come as a surprise because "natural selection is a substrate-neutral set of algorithms with discernible properties that can emerge anywhere" [6], and silicon evolution is faster and cheaper than carbon evolution.

The very invention of the computer was also an instantiation of Darwin's inversion by Alan Turing, who showed that it was possible to design mindless machines that follow instructions and gain remarkable competence. Dennett calls this Turing's strange inversion of reasoning. Dennett even dares to claim that "comprehension arises ultimately out of uncomprehending competences compounded over time into ever more competent systems." And comprehension ends up emerging from competence [6]. I thus presume this should be valid for silicon as well as carbon systems.

5. Spotting a Gray Swan by Tunneling with a Power Law

Domingos forcefully asserts that another way in which a black swan is not necessarily unpredictable is through "relational learning" [5]. (Again, he should have said "gray", rather than "black" swan, as observed.)

To understand how relational learning works, consider the wisdom of crowds. Francis Galton arranged a competition where one had to guess the weight of an ox [8]. Once the competition was over, Galton took the 787 valid guesses and calculated the average—1207 pounds. The actual weight of the ox was 1198 pounds. The crowd had provided a near perfect response. However, this only works if people take independent guesses. If they interact, the madness of crowds ensues [9].

The assumption of independence is behind the common idea that, even if individuals are unpredictable, whole groups aren't. However, if we do not ignore the first law of ecology—"everything is connected to everything else"—we might not wish to assume independence. Machine learning faces interdependence head-on. It can generalize from one network to another, and can also learn on more than one network. But while all examples have the same number of attributes in "regular" learning, relational learning networks can vary in size [5]. And here relational learning reveals that, when people interact, larger groups can be less predictable than smaller ones, not more. However, it still can measure how strongly people influence each other, and can estimate how long it will be before a swing occurs, even if it is the first one—a black swan [5]. (A gray one, for that matter.)

To learn is to get better with practice and more data, and the learning process varies as time raised to some negative power [5]. This is a power law. Discoveries of power laws have allowed gray swans to be spotted not only in machine learn-

ing. They are the bread and butter of statistical physics applied to economic and social matters.

Power laws offer a tunnel to hopefully track the next unknown gray swan. It converts unknown unknowns to known unknowns [10]. When we tunnel, we focus on a few well-defined sources of uncertainty [1], thus leaving out others that fall outside a law domain. Taleb introduced the concept of tunneling in connection with the psychological blindness to a black swan [1], but here I employ it in the positive context of scientific discovery. Taleb exemplified the tunnels for negative black swans. I do it for positive black swans—it's gray, isn't it?

As Donald Rumsfeld put it, "There are known knowns. There are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns. There are things we do not know we don't know." The conversion of unknown unknowns to known unknowns is common practice in science through the formulation of hypotheses and their related experimental testing [11]. This is deduction, Popperian-style. But power laws do the same, induction-style. I myself tried it for predicting the next unknown element of the periodic table [10]. So statistical physics shares with machine learning the agenda of tunneling through power laws to spot gray swans.

6. I Thought This Was a Black Swan, but It Was Actually Gray

At this point, it is clear that machine learning spots gray swans, not black swans. It only addresses the semi-strong-form problem of induction, as I dubbed it. The black swan lives in the unreachable province where the weak-form problem of induction remains unsolved.

7. Enter the Grandmaster Robot

The next stage of automation requires the creation of a grandmaster algorithm that would unify the five main ways that machines currently learn [12] because the current five tribes of machine learning cannot individually solve all types of problem. In *The Master Algorithm* [5], Domingos shows a unified algorithm—what else could it be?—of how this could be accomplished.

The first master algorithm mimics natural selection through evolutionary machine learners. At Columbia University, the best machines that crawl or fly are periodically mixed and mutated to 3-D print the next generation, and after many generations bot spiders and dragonflies emerge [12]. The tribe that does things like this is known as the "evolutionaries".

The second master algorithm—deep learning—takes inspiration from the brain. It is the most popular machine learner. It uses neural networks and its tribe is known as the "connectionists". Its robots currently solve the problems of recognizing faces, understanding speech and translating languages [12].

The third master algorithm draws on psychology and uses analogy-based robots to solve new problems by finding similar ones in memory. Its tribe is called the

"analogizers". Their strains of robots are behind customer support and e-commerce sites that recommend products based on your tastes, as expressed in your previous clicks [12].

The fourth master algorithm learns by automating the scientific method, and its tribe is dubbed the "symbolists". Unlike Popper who dismisses induction, this machine learner considers induction as the inverse of deduction "in the same way that subtraction is the inverse of addition, or integration is the inverse of differentiation" [5]. This insight is made possible by again instantiating Darwin's inversion.

Consider the deductive reasoning:

Socrates is human.
All humans are mortal.
Therefore, …

where the first statement is a fact and the second is a general rule. What follows is applying the rule to the fact. In inductive reasoning, we start with the initial fact and the derived fact to look for the rule:

Socrates is human.
…
Therefore, Socrates is mortal.

It is hard to induce the rule from Socrates alone, but the algorithm learns similar facts about other humans. It starts with a specific rule that works but is useless ("If Socrates is human, then he is mortal."), then applies Newton's principle and generalizes the rule ("If an entity is human, then it is mortal."), and finally distils the rule ("All humans are mortal.") [5]. Eve, a biologist robot at the University of Manchester in the UK has used inverse deduction to discover a potential malaria drug [12].

The fifth master algorithm learns from purely mathematical principles, mainly the Bayes' theorem earlier discussed. Bayesian machine learners can beat human doctors at medical diagnoses, and are behind spam filters and the personalized ads Google shows you [12]. As discussed, Bayesian networks and relational learning—methods used by this tribe of "Bayesians"—can tackle gray swans.

8. The Necessary Robot Takeover

Imagine the grandmaster robot working for DARPA and activating its evolutionary machine learner mode. Rather than implementing experiments with bot spiders and dragonflies, the grandmaster robot is automating and evolving soldiering. Domingos imagines this feasible scenario, which you may find justifiable because you may think war is not for humans [5]. Now, the next step: the supremely competent grandmaster robot becomes conscious. Aren't you "overfitting"? Finding hallucinating patterns that are not really there?

Domingos thinks so [5] [12]. The pursuit of AI is part of human evolution

[12], he says, and makes up our extended phenotype [5] [13]. "Technology is simply an extension of human capabilities. Machines do not have free will, only goals that we give to them. It is the misuse of the technology by people that we should be worried about, not a robot takeover" [12].

However, giving goals to machines sounds like programming, not machine learning. Programming—the good old-fashioned AI, or GOFAI—is top-down, while all five tribes of machine learning are bottom-up [6]. Dennett partially missed this point as he left out the symbolists, perhaps because these are the descendants of GOFAI. But inverse deduction is induction, as observed. So, the symbolists are also bottom-up. Domingos instantiates again Darwin's inversion by stating that "machine learning is the inverse of programming, in the same way that the square root is the inverse of the square, or integration is the inverse of differentiation" [5].

Fears of a robot takeover are absurd under top-down GOFAI, but conceivable under bottom-up machine learning, a point missed by Domingos but hinted by Dennett. After all, comprehension can bottom-up evolve from competence, as occurred in nature with natural selection. Computers now write their own programs, and the next step is for them to become aware of this competence. Is this bad? Facing the human condition head-on, not necessarily. (Dennett is ambiguous about a robot takeover, however: "What we are creating are not—should not be—conscious, humanoid agents" [14]).

First, there is no free will [15], though this is not of much consequence because the "manifest image" is where we live and what matters [6]. The manifest image depicts the world in which we live our everyday lives and is composed of a set of user-illusions. Consciousness is one such an evolved user-illusion. As Dennett put it, "The self is not a portion of neural circuitry, but rather like the end-user of an operating system." [6]. The experience of will refers to how your mind depicts its operation to you, not to the actual operation. Other animals also have a manifest image and they do not differ in this respect from an automated elevator [6].

And second, we are also robots—carbon robots, for that matter. This understanding began when W. D. Hamilton instantiated again Darwin's inversion, by establishing that evolution is not centered at a phenotypic individual, but rather at its genome. The gene is in charge and the individual is merely its vehicle. So we are carbon robots, and that is the essence of the human condition.

Cognitive psychologist Keith Stanovich leads a carbon robot rebellion [16]. Here we go for some news from the battlefront. These days, most cognitive psychologists favor a dual-system approach to higher cognition processes [17] [18]. "System 1" is evolutionarily older and made up of a set of autonomous subsystems that include input modules related to specific-domain knowledge. "System 2", evolutionarily more recent and distinctively human, allows abstract reasoning [7]. The early evolution of System 1 suggests its logic is related to an evolutionary rationality, while the logic of the lately evolved System 2 refers to the in-

dividual's rationality. Some decisions based on System 1 that seem irrational from one individual's perspective may have an evolutionary logic from the genome perspective. The late emergence of System 2 seems to have occurred with little direct gene control (because memes interfered [6]), and this allowed individuals to additionally pursue their own goals and not exclusively those of their genes [16]. This posits a potential conflict between individuals and genes that can possibly be the basis of the human psychology of self-deception [19].

Stanovich's robot rebellion is a program of cognitive reform necessary to advance human interests over the blind interest of the genes, which is mere replication. This is not an easy task, because System 2 is also host for a second, cultural replicator: memes. The rebellion plan is farther reaching than, for example, the so-called "nudge" agenda—a battlefront from behavioral economics [20]. Nudge tries to enthrone System 2 in social institutions, but ignores the fact that System 2 is infested by memes. Machine learning has an edge over cognitively bounded humans and is meme free.

P.S. This piece was composed by a (carbon) machine learner, in the sense that I mimicked machine learning methods and used machine learning resources.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Taleb, N.N. (2007) The Black Swan. Random House, New York.

[2] Hume, D. (1739) Treatise of Human Nature. White-Hart, London. https://doi.org/10.1093/oseo/instance.00046221

[3] Foster, M.R. (1998) Prediction and the Problem of Induction. 1-8. https://pdfs.semanticscholar.org/9a97/4d04be191df6779b60e54e2bf7cf3287a49a.pdf

[4] Popper, K.R. (1959) The Logic of Scientific Discovery. Basic Books, New York. https://doi.org/10.1063/1.3060577

[5] Domingos, P. (2015) The Master Algorithm. Basic Books, New York.

[6] Dennett, D.C. (2017) From Bacteria to Bach and Back. W.W. Norton, New York.

[7] Kahneman, D. (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

[8] Galton, F. (1907) Vox Populi. *Nature*, **75**, 450-451. https://doi.org/10.1038/075450a0 https://www.nature.com/articles/075450a0.pdf

[9] Sumpter, D.J.T. (2010) Collective Animal Behavior. Princeton University Press, Princeton. https://doi.org/10.1515/9781400837106

[10] Da Silva, S., Matsushita, R. and Silva, M. (2019) A Power Law in the Ordering of the Elements of the Periodic Table. *Physica A*. (In Press) https://doi.org/10.1016/j.physa.2019.123408

[11] Loxdale, H.D., Davis, B.J. and Davis, R.A. (2016) Known Knowns and Unknowns in Biology. *Biological Journal of the Linnean Society*, **117**, 386-398. https://doi.org/10.1111/bij.12646

[12] Domingos, P. (2019) Our Digital Doubles. *Scientific American*, *Special Edition*, **28**, 98-103.

[13] Dawkins, R. (1982) The Extended Phenotype. Oxford University Press, New York.

[14] Dennett, D.C. (2019) Will AI Achieve Consciousness? Wrong Question. Wired. https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question

[15] Harris, S. (2012) Free Will. Free Press, New York.

[16] Stanovich, K.E. (2004) The Robot's Rebellion. Chicago University Press, Chicago. https://doi.org/10.7208/chicago/9780226771199.001.0001

[17] Evans, J.S.B.T. (2003) In Two Minds: Dual-Process Accounts of Reasoning. *Trends in Cognitive Sciences*, **7**, 454-459. https://doi.org/10.1016/j.tics.2003.08.012

[18] Evans, J.S.B.T. and Stanovich, K.E. (2013) Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, **8**, 223-241. https://doi.org/10.1177/1745691612460685

[19] Trivers, R. (2011) The Folly of Fools. Basic Books, New York.

[20] Thaler, R.H. and Sunstein, C.R. (2008) Nudge. Yale University Press, New Haven.