Training Data and Rationality

Konstantinos Mersinas[1], Theresa Sobb[2], Char Sample[3], Jonathan Z. Bakdash[4], and David Ormrod[5]
[1]Royal Holloway University of London, London, UK, Konstantinos.Mersinas@rhul.ac.uk
[2]University of New South Wales at Australian Defence Force Academy, Canberra, AU, t.sobb@student.unsw.edu.au
[3]ICF International, Fairfax, VA, US, Char.Sample@icf.com
[4]U.S. Army Combat Capabilities Development Command – Army Research Laboratory South at the University of Texas at Dallas, Richardson, TX, US, jonathan.z.bakdash.civ@mail.mil
[5]Campbell, AU, drdave@linux.com

Abstract

Human decision-making includes emotions, biases, heuristics within environmental context, and does not generally comply with rational decision-making (e.g. utility maximization). Artificial Intelligence (AI) algorithms rely on training data for analysis, based upon observed cybersecurity incidents. Rarity of attack data and the existence of information asymmetries between attackers and defenders creates uncertainty in estimating attack frequencies and can reduce the reliability of data. These characteristics can lead to a posteriori justification of attacker/defender choices, deeming successful actions as rational, and vice versa.

Data analysis also influences the fidelity of AI output. The need for broad specification analysis often leads to analysing large amounts of data and using increasingly complex models. The fuzzy definition of rationality creates an opening for exploitation. Data volumes and model complexity may consequently reduce the usefulness of predictions in this fuzzy environment.

AI lacks adaptive human characteristics like “common sense” that make humans situationally adaptive. AI relies on learning and prioritizing likely events; however, lower probability ‘common sense’ defying can, upon success, reorder probabilistic outcomes (Gersham, Horovitz and Tenenbaum 2015). Such characteristics are not easily quantifiable in AI environments and can therefore decrease the efficacy of machine learning (ML) classification processes. This research effort differs from traditional Adversarial ML goals; we expose inherent flaws in “good” historical data, based on successful lessons learned. In such cases, “rational” ML optimization potentially produces misleading results, often unlikely to trick humans, reflecting the potential limitations of AI and pattern recognition given the prior and available data. The exposure is cognitive not algorithmic.

Considering the aforementioned factors, propagated biases and “irrational” choices, if combined with issues inherent in data analysis and current AI capability limitations, weaken the predictive power of AI. These problems provide reason to consider ”rationality” through the lens of AI-reliant cybersecurity, potentially weakening security posture. Psychological aspects heavily influence the way humans approach, understand, and act to solve problems. Consequently, human-originated historical data in cybersecurity may be of reduced utility for AI, due to a lack of contextual information.

This paper provides an introductory overview and review of AI in the context of human decision making and cybersecurity. We investigate the notion of “rationality” and types of AI approaches in cybersecurity, discussing the differences between human and AI decision-making. We identify potential conflicts between human decision-making biases and AI data analysis.

Keywords: *artificial intelligence, data, decision-making, rational, rationality, bias*

1. Introduction

Human decision-making differs from artificial intelligence (AI) in several critical areas (Lake et al. 2017). Human decisions are influenced by values, beliefs and biases, all of which may not be rational. Despite the imbalance between logic and emotions, there have been times when illogical choices have, in retrospect and via success, been observed as correct and in some cases as outstanding choices. Some historic disruptive decisions that were only in retrospect considered the right choice include Henry Ford’s 1914 controversial

decision to double the salaries of his workers (Fortune 2012); a decision considered irrational at the time since high labour costs cut into corporate profits. However, Ford recognized that by paying workers more they could afford the cars they were producing; significantly increasing the product's demand (Ibid). Similarly, the return of Steve Jobs in 1997, to Apple was considered a mistake since he had left Apple in 1985 under non-favourable terms (Fortune 2012; Terdiman 2013). Both cases illustrate changes to rules determining rational decisions that became the new standard "rational" decision.

Now consider whether a machine would make different decisions. In both cases the opposite choice would have prevailed. Reinforcement learning would be the mechanism to correct this oversight, but only in retrospect. The use of reinforcement training in cybersecurity is reactive and suggests a flaw in the original training data. Furthermore, a question about the efficiency of re-classifying data arises.

The problem being addressed in this paper is the potential for AI algorithms to perform as expected, yet yield incorrect or suboptimal results. AI-enabled cybersecurity responses require proactive measures; failure to do so results in exposure and potential exploitation. The present state of AI-enabled cybersecurity is a faster reactive model that risks leaving assets exposed while providing the impression of proactive security. To support this analysis, this paper provides an overview of human decision making and the implications for AI as it applies to cybersecurity.

2. Background

Human decision-making is a complex process and one that is not necessarily rational. Decisions are shaped by emotions (LeDoux 1998), biases and heuristics (Kahneman 2011), and the environment/context (Gigerenzer and Selten 2002; Hutchins 1995; Norman 2013). Thus, human decision-making generally does not follow rational decision-making (e.g. profit/utility maximization); Kahneman (2011) provided supporting evidence for the irrational perspective. Even when humans believe they are making a conscious decision, the ratio of unconscious to conscious data overwhelms the person, making a fully-conscious decision highly unlikely (Dijksterhuis 2004; Bargh and Morsella 2008; Evans 2008). Despite the irrational nature of decision-making, humans have made rational decisions throughout the course of history, even if some of those decisions were considered rational only in retrospect.

The neuroplasticity of the human mind associates with the ability to switch mental algorithms and has not yet been fully replicated by AI (Wagarachchi and Karunananda 2017). Neural networks recreate plasticity traits, like data pruning, associated with the human brain, the environmental variables that require monitoring have not been fully enumerated (Williamson 2014). Thus, while aspects like pruning or prioritizing can occur, the selection of environmental variables remains a work in progress (Schoettle and Sivak 2015).

2.1 Rationality Types

We briefly present some well-known theories and approaches to 'rationality'. Of note, these approaches are not (always) mutually exclusive, and many overlap in various ways. Moreover, there is no single, universal theory to explain and predict all human decision-making. We also recognize an existing adaptive component to rationality, which separates human- from artificial intelligence (Gershman et al. 2015).

2.1.1 Full (unbounded) Rationality and Rationality as optimisation

Expected Utility Theory

Expected utility theory assumes substantively rational expected utility maximizers, where utility can be subjective. Individual rationality is a self-aware process that calculates the maximization of gains (or loss minimization). The formal version of expected utility theory proposes a set of axioms (von Neumann and Morgenstern 2007). Preferences of individuals can be ranked via a utility function, and the assertion is that rational individuals comply with the rationality axioms in order to maximize their expected utility.

The quantitative optimisation approach for deciding cyber security investment levels is usually studied in the spirit of Gordon and Loeb (2002). Given a vulnerability, the level of investment is decided by a cost-benefit

analysis on the total investment expenditure and the expected losses from invoking that vulnerability. The approach includes a number of estimation trials.

Game theoretic approaches also follow the normative paradigm of full rationality and optimisation. For example, two-stage games have a typical setting of an attacker and a defender deciding on attacks and security controls simultaneously (Hogarth and Kunreuther 1995). Such models provide useful insights, but often include non-practical assumptions, such as the full elimination of vulnerabilities.

Substantive and Procedural rationality

Whether an action is substantively rational depends only on decision-makers' goals. Behaviour is appropriate if it is meant to achieve a specific goal (including subjective utility) under specified conditions. According to procedural rationality, behaviour can be the outcome of an appropriate deliberation process, or the outcome of impulsive, affective mechanisms and without the intervention of reason. In the former case decisions are labelled as rational, and the latter as irrational.

2.1.2 Types of Rationality and Non-Rationality

Ecological and intuitive rationality

Types of rationality can be summarized as bounded rationality. These limitations can be cognitive, like mental capacity, or the lack of information or time for a decision. Ecological rationality targets behaviours which exploit environment information structures through a number of heuristics, instead of utility maximization (Gigerenzer and Selten 2002; Goldstein and Gigerenzer 2002). Ecological rationality suggests that observed behaviour needs to be examined against contextual conditions, not against normative rules. In that sense, ecological rationality can emerge from the interaction of the environment and the heuristics that decision-makers use. The environment may have been structured by institutions which have different incentives than those of the individual (Gigerenzer 2008).

Intuitive rationality resembles pattern-matching, but takes place at an unconscious level (Simon, 1965). "Results" or specific choices emerge fast and effortlessly to consciousness, but the process remains hidden from the individual. Immediate and intuitive responses to problems are linked with proficiency and expertise (Zsambok and Klein 1997). We examine intuition in more detail in Section 4.3.

2.1.3 Heuristics and Biases

In sharp contrast to Expected Utility Theory, Kahneman (2011) found widespread evidence of heuristics and biases in human decision-making. That is, violations of rationality axioms phenomena such as loss aversion due to a framing effect (see Section 4). Kahneman (2011) specifies two systems for decision-making: System 1 (fast and automatic) and System 2 (slower, deliberate and consciously strenuous). System 1 shares similarities with heuristics in ecological rationality.

2.2 Types of AI

The general approach to AI has been the development of a fully rational agent, according to unbounded rationality and the optimisation paradigm, based on an exhaustive cost-benefit analysis. There are exceptions to this approach, such as learning reward functions (see Section 3). The unbounded rationality approach is justified by the fact that AI does not have the limitations in time and computational power that humans exhibit. Humans have evolved to move between various rationality models seamlessly as the situation/environment demands, with limited memory capacity and time. Full-search 'brute force' approaches are not necessarily suitable for inductive ML, as they often contain significant noise and miss feature values (Zhang, 2002).

2.3 Historical data issue of small sample statistics

Historical cybersecurity data constitutes small and not-necessarily representative samples, allowing for the phenomenon of regression to the mean and leading to misinterpreted predictions of future events (Brighton and Gigerenzer 2015). If an initial measurement is extreme, then subsequent measurements are likely to be closer to the mean. Regression to the mean occurs whenever correlation between two scores is imperfect (Kahneman 2011). Successful attacks depend on skills, conditions and luck. 'Extremely' damaging attacks are characterised according to the unknown distribution of attack outcomes, and future attacks are likely to be less damaging, as they will be less extreme.

3. AI and Machine Learning Techniques

AI refers to the ability of a machine, system or network to exhibit "human intelligent behaviours", including sensing, perception, reasoning, thinking and learning (Li and Du 2017). Machine learning (ML) exists within the AI sphere, focusing on designing and operating computers which can improve over time, and determining the "fundamental statistical-computational-information-theoretic laws" that apply to these systems (Jordan and Mitchell 2015). We define interpretability of AI/ML as "the degree to which an observer can understand the cause of decisions", using the terms explanation and interpretability interchangeably (Miller 2017). ML uses algorithms to achieve descriptive, predictive and prescriptive performance enhancement in realistic scenarios, attempting to optimise performance using provided or past-experience data (Alpaydin 2009; Cochran 2018). Figure 1 describes the ML structure, where the learner uses inputted information to create a model, which can then be used to make predictions relating to new data provided to the system.

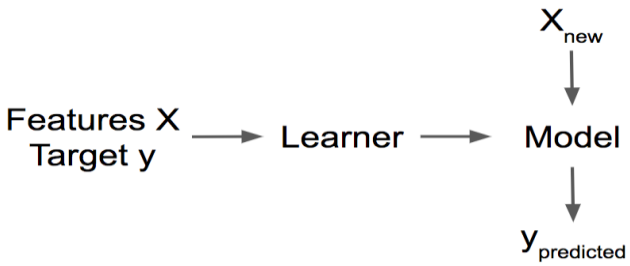

Figure 1: Machine Learning Model (Molnar 2018)

The breadth of problems addressed by ML algorithms spans from data mining to gaming (Kaelbling, Littman and Moore 2006; Kotsiantis, Zaharakis and Pintelas 2007). A variety of ML categories have been created to define how algorithms meet different problem and solution needs; including supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning (Ayodele 2010; Fumo 2017).

ML relies on the identification of patterns to understand processes and make predictions, with models utilised including probability distributions, linear regression models, linear classification models, neural networks, kernel methods, sparse kernel machines, graphical models and mixture models, among others (Alpaydin 2009; Bishop 2006). ML can be categorised as supervised, unsupervised, semi-supervised or reinforcement learning.

Supervised learning uses an example input labelled dataset and associates features of the data inputs with the output labels (Lison 2015). Classification and regression models are forms of supervised learning (Zhu and Goldberg 2009). Algorithms include nearest neighbour, naive bayes, decision trees, linear regression, support vector machines and neural networks (Fumo 2017; Kotsiantis, Zaharakis and Pintelas 2007). An example of supervised classification is the filtering of email as spam (Renuka et al. 2011). An example of supervised regression is modelling the connection between age and dementia (Zhu and Goldberg 2009). Unsupervised learning involves a range of data inputs with no output labels, creating outputs based on their features (Lison 2015). Clustering and outlier detection models are examples of unsupervised learning (Zhu and Goldberg 2009). Algorithms include k-means clustering and association rules (Fumo 2017; Kavakiotis et al. 2017). An example of unsupervised clustering includes the identification of people based on their individual gaits (Ball 2012). An example of unsupervised outlier detection includes anomaly detection (Eskin et al. 2002).

Semi-supervised learning includes a combination of supervised and unsupervised learning techniques (Zhu and Goldberg 2009). Some, but not all, of the data is labelled; and the machine uses these available parameters to determine groupings (Fumo 2017; Zhu and Goldberg 2009). Semi-supervised learning models include self-training, mixture models and semi-supervised support vector machines (Zhu and Goldberg 2009).

Reinforcement learning machines interact with the environment, receiving rewards or penalties for actions in hindsight, and are taught the actions needed to optimise rewards through trial and error (Alpaydin 2009). The technique is sometimes called "learning with a critic" (Alpaydin 2009). Q-Learning and Temporal Difference algorithms are used in reinforcement learning (Fumo 2017; Kaelbling, Littman and Moore 2006).

4. Human Decision-Making and AI

Human and AI decision making differ in several ways. For example, the use of logically equivalent frames, intuition, preferences, motivations, context and biases can all affect human decision making differently to AI.

Logically equivalent frames

Humans unconsciously contextualize information, often in the form of biases. However, contextualization can also signify expanded perception (Brighton and Gigerenzer 2015). Experiments demonstrate instances where participants have assigned intent and human attributes to geometric shapes and objects, based on the way they "interacted" with each other (Heider and Simel 1944). These experimental characterisations are fundamentally human concepts. ML techniques which lack labelled data, such as unsupervised learning, would be unable to classify the shapes in this way. Should a computer group these shapes like human respondents did, they would require the provision of information on human behavioural roles.
Additionally, there is a level of "logical abstraction" when decisions of attackers/defenders are analysed by AI, that might omit cues which humans have possibly taken into consideration before making decisions.

Intuition, unconscious and experience

If intuition is defined as a rational, but unconscious, pattern recognition (Simon, 1987), then it could be mimicked by AI. Heuristics are an example of AI replicating the human unconscious thoughts or intuition. However, humans often have trouble accepting or justifying intuitions. Skilled experience in a subject or task is one of the main sources of useful intuitions, especially with regards to identifying cues in an environment. Environments need to allow for causal and statistical properties to manifest and provide decision makers with opportunities to learn important cues and develop skills. In the absence of skills, people are susceptible to incorrect intuitions, due to cognitive biases, like memory operations and anchoring effects (Kahneman and Klein, 2009).

Accidents have revealed that autonomous vehicles have misunderstood situations and humans expected that other drivers would understand and interpret their intentions correctly, which was not the case with the autonomous vehicles AI (Anthony, 2017). Humans perceive action and intention, with many road rules containing exceptions based on the sensory feedback loop that informs common sense. Self-driving cars may lack the social or situational awareness to understand this concept, thus making sub-optimal decisions. Common sense and intuitively interpretable cues which are natural for humans, even if conditioned under skill, are far from easy for AI.

Preferences

Preferences can also introduce biases by influencing how people see the future. Perceived probabilities are adjusted by individuals according to subjective preferences, and thus, the likelihood of outcomes and appropriate actions might lead to subjectively diversified decisions, as in the overconfidence bias (DeBondt and Thaler 1995). Likelihood estimations, especially of rare events, depend on whether a decision is based on experience (rare events underestimation) or from description of problems (rare events overestimation) (Hertwig and Erev 2009). Machines are not directly affected by such distortions, through their objectiveness. However, human preferences can indirectly induce biases, as supervised, semi-supervised and reinforcement

learning ML techniques rely on some form of human input (Alpaydin 2009; Lison 2015; Zhu and Goldberg 2009).

Motivations

There are difficulties with analysing human motivation and dedication. For AI to achieve a human-like intelligence, it should manifest open-ended learning, progressing to increasingly more complex skills, driven by intrinsic motivations for learning these skills and obtaining knowledge (Baldassarre, 2011). The breadth of human intrinsic motivations constitutes a hard-to-identify (and hard-to-mimic) factor for AI. Understanding how attackers select targets is an open problem (Kenneally et al., 2018) and could be an important complementary factor in attacker profiling approaches; such as zero-day attack profiling (Sample et al. 2017). Indicatively, asset values cannot be aligned with espionage or national state agent profiles, because such attackers assign greater values to targeted assets than organisations do (Lakhani and Wolf, 2003). Motivations and attackers' "personal agendas" are not revealed in AI analysis and, consequently, the evaluation of attacks can be skewed.

Context, information asymmetries and decision outputs

Humans employ heuristics to make decisions. The effectiveness of simple heuristics in various areas has been shown to depend on the context/environment in which they are made, e.g. the size of the dataset (Katsikopoulos et al. 2010). Contextual factors and the process that leads to decisions are often underlying and hidden, thus, only the final decision is observable. Consequently, potentially important factors are not recognised by AI.

Similarly, information asymmetries imply that during a decision involving two or more agents, one of the agents has access to information that others do not have. Machines that monitor and analyse the outcome at the end of a causal chain will generally be agnostic to the decision processes and the contextual factors which lead to the decision. A human conducting an analysis, however, would be more likely to consider possible environmental and procedural factors leading to the decision.

Decision Making Biases and Heuristics

People estimate the probability of an event based on its similarity with its parent population, according to the *representativeness* bias. *Availability* inflates the likelihood of salient events in the individual's memory. By *recognition* people consider recognisable events/objects as more likely to happen. Decision makers' predictions are unconsciously correlated with initial "anchors" of irrelevant information (Kahneman 2011; Gigerenzer and Gaissmeier, 2011).

For example, a social environment can be a group of hackers with members behaving by drives for status elevation or name-building and acceptance amongst peers or even for a sense of accomplishment. There could also be behaviours motivated by imitation of peers. Such factors insert complexity in the decision-making process and this complexity is not mapped in AI analysis.

The *satisficing* heuristic has individuals search amongst alternative actions until an alternative is found, which satisfies a relevant aspirational level. The alternative that is selected needs to be "good enough" to satisfy the aspiration level but is not necessarily optimal (Simon 1986).

Heuristics could be considered a human advantage, as they reduce the need for cognitive resources and time. Unsupervised learning algorithms can mimic non-utility maximising behaviour by clustering, where the goal is to find similarities in training data (e.g. Amazon's book recommendation algorithms); however, sufficient amounts of data are required and the problem of overfitting the training data is ever present (Ayodele 2010).

Human decisions can be influenced by *social factors* like group loyalty, transparency and accountability. Social contexts allow for imitation behaviours, reciprocation actions, exploiting the "wisdom of the crowds" by combining or averaging opinions of others, and being influenced by social circles (Hertwig and Herzog, 2009).

Finally, humans are susceptible to variations of a problem's presentation, which can cause preference reversals and decision changes, even amongst experts (Mersinas, 2015).
Due to logical abstractions of problem descriptions AI is hardly susceptible to *framing effects*. However, AI would not be able to explain the existence of two contradicting decisions of a human in seemingly identical conditions, because it would not have the ability to understand the different frames under which the two decisions were made.

5. Discussion and Conclusion

Overall, the design of AI has left psychological aspects of decision-making outside its scope. This approach reflects the full rationality and optimisation paradigm, in the same way that economics was "freed from any dependence upon psychology", due to the assumptions of utility maximization and substantive rationality (Simon 1976). However, psychological aspects heavily influence the way humans approach, understand, and act to solve problems. Consequently, the utility of human-originating historical data in cybersecurity is potentially reduced for AI.

One challenge with using human supplied data in AI comes from misalignment between systems. AI, while replicating biological principles of intelligence, is rarely designed to be an artificial manifestation of human brains, but rather serve to solve specific problems or to respond within a set scenario. An inevitable compromise will occur between having highly specialised machines in their problem environment, and machines that understand the complexities of cause and effect from a human conceptual perspective. This trade-off can be considered as being between the value of contextual understanding (understanding of the human factors which may influence a decision) and the computational costs of gaining that understanding and responding to it (Gershman, Horvitz and Tenenbaum 2015).

Most ML techniques rely on some form of human input or labelling, and thus their decisions can become subject to human flaws, such as subjective perception and biases. Decisions are driven by the various forms of heuristics and "rationality types" which humans naturally use, and which cannot be identified or decoded by AI. Even in cases of minimal human intervention in the algorithms' learning, there are inherent data issues, such as, noise removal, data over-fitting, and lack of inductive learning efficacy. Today's AI mechanisms have limitations. For AI to be able to mimic and understand human behaviours, AI must evolve into obtaining intrinsic motivations and curiosity for learning and obtaining knowledge. Furthermore, AI needs to develop mechanisms for automatically switching between heuristic (unconscious) and computational (conscious) algorithms based on environmental (contextual) factors. At a more practical level, historical cybersecurity data requires deeper-level analysis to be realistically understood. Sufficiency of data, rare events and misunderstood regressions to the mean are all factors that diminish data fidelity. For decisions analysed by AI algorithms, the context of decision-making, the psychological state of the actors, and the type of rationality that the actors utilised for their choices, all constitute hindering factors for AI to understand and interpret historical cybersecurity data effectively.

This paper provides an introductory overview and review of AI in the context of human decision making and cybersecurity. The issue of human decision-making biases and the practical development of cybersecurity AI solutions remains an unresolved problem space, for further research.

6. Acknowledgements

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

The views expressed are the authors' and not necessarily those of the Royal Australian Air Force, Australian Army or the Department of Defence. The Commonwealth of Australia will not be legally responsible in contract, tort or otherwise for any statement made in this publication.

7. References

Alpaydin, E., 2009. *Introduction to machine learning*. MIT press.

Anthony, S. 'Self-driving cars still can't mimic the most natural human behaviour', 2017, accessed 25/5/2019, <https://qz.com/1064004/self-driving-cars-still-cant-mimic-the-most-natural-human-behavior/>

Ayodele, T.O., 2010. Types of machine learning algorithms. In *New advances in machine learning*. Arel, I., Rose, D.C. and Karnowski, T.P., 2010. Deep machine learning-a new frontier in artificial intelligence research. *IEEE computational intelligence magazine, 5*(4), pp.13-18.

Baldassarre, G., 2011. What are intrinsic motivations? A biological perspective. *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics*.

Ball, A., Rye, D., Ramos, F. and Velonaki, M., 2012, March. Unsupervised clustering of people from 'skeleton' data. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp.225-226). IEEE.

Bishop, C.M., 2006. *Pattern recognition and machine learning*. Springer.

Brighton, H. and Gigerenzer, G., 2015. The bias bias. *Journal of Business Research*, *68*(8), pp.1772-1784.

Cochran, J.J.J., 2018. *Informs analytics body of knowledge*, Wiley.

DeBondt, W. F. and R. Thaler. 1995. "Financial decision-making in markets and firms: A behavioral perspective". In: Handbook in Operations Research and Management Science, Vol. 9, Finance. Ed. by R. A. Jarrow, V. Maksimovic, and V. T. Ziemba. North Holland: Elsevier. 385–410.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L. and Stolfo, S., 2002. A geometric framework for unsupervised anomaly detection. *Applications of data mining in computer security* (pp.77-101). Springer, Boston, MA.

Fumo, D. 2017, *Types of machine learning algorithms you should know*, Towards Data Science, accessed 17/05/2019, <https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861>.

Gershman, S.J., Horvitz, E.J. and Tenenbaum, J.B., 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), pp.273-278.

Gigerenzer, G., 2008. Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20-29.

Gigerenzer, G. and Gaissmaier, W., 2011. Heuristic decision making. *Annual review of psychology*, *62*, pp.451-482.

Gigerenzer, G., Selten, R. (Eds.), 2002. *Bounded rationality: The adaptive toolbox*. MIT press.

Goldstein, D. G., Gigerenzer, G., 2002. Models of ecological rationality: the recognition heuristic. *Psychological Review*, *109*(1), 75-90.

Gordon, L.A., Loeb, M.P., 2002. 'The economics of information security investment'. *ACM Transactions on Information and System Security (TISSEC)*, 5(4), 438-457.

Heal, G., Kunreuther, H. (2005) 'IDS models of airline security'. *Journal of Conflict Resolution*, 49(2), 201-217.

Heider, F. and Simmel, M., 1944. An experimental study of apparent behaviour. *The American journal of psychology*, *57*(2), pp.243-259.

Hertwig, R. and Erev, I., 2009. The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), pp.517-523.

Herzog, S.M. and Hertwig, R., 2009. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), pp.231-237.

Hogarth, R.M., Kunreuther, H. (1995), 'Decision making under ignorance: Arguing with yourself'. *Journal of Risk and Uncertainty*, 10(1), 15-36.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), pp.255-260.

Kaelbling, L.P., Littman, M.L. and Moore, A.W., 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research, 4*, pp.237-285.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Macmillan.

Kahneman, D., Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologis*t, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Katsikopoulos, K.V., Schooler, L.J. and Hertwig, R., 2010. The robust beauty of ordinary information. *Psychological Review*, *117*(4), p.1259.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal, 15*, pp.104-116.

Kenneally, E., Randazzese, L. and Balenson, D., 2018, June. Cyber Risk Economics Capability Gaps Research Strategy. In 2018 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)(pp.1-6). IEEE.

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, pp.3-24.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B. and Gershman, S.J., 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

LeDoux, J. (1998). *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster.

Li, D. and Du, Y., 2017. *Artificial intelligence with uncertainty*. CRC press.

Lison, P., 2015. An introduction to machine learning.

Morgenstern, O. and Von Neumann, J., 1953. *Theory of games and economic behaviour*. Princeton university press.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, pp.1-38

Molnar, C., 2018. Interpretable machine learning: A guide for making black box models explainable. *Christoph Molnar, Leanpub*.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

Renuka, D.K., Hamsapriya, T., Chakkaravarthi, M.R. and Surya, P.L., 2011. Spam classification based on supervised learning using machine learning techniques. *International Conference on Process Automation, Control and Computing* (pp.1-7). IEEE.

Sample, C., Cowley, J. and Hutchinson, S., 2017. Cultural exploration of attack vector preferences for self-identified attackers. *11th International Conference on Research Challenges in Information Science (RCIS)* (pp. 305-314). IEEE.

Simon, H.A., 1965, Administrative behaviour (2nded). New York:Free Press.

Simon, H.A., 1976. From substantive to procedural rationality; 25 years of economic theory (pp.65-86). Springer, Boston, MA. Simon, H.A., 1986. Rationality in psychology and economics. *Journal of Business*, pp.S209-S224.

Zhang, S.Z., 2002. Data Preparation for Data Mining. *Applied Artificial Intelligence*. 17,375-381.

Zhu, X. and Goldberg, A.B., 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning, 3*(1), pp.1-130.