Toward out-of-distribution generalization through inductive biases

Citation for published version:
Moruzzi, C 2022, Toward out-of-distribution generalization through inductive biases. in VC Müller (ed.), *Philosophy and Theory of Artificial Intelligence 2021.* Studies in Applied Philosophy, Epistemology and Rational Ethics, Springer, pp. 57-66. https://doi.org/10.1007/978-3-031-09153-7_5

Digital Object Identifier (DOI):
10.1007/978-3-031-09153-7_5

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Philosophy and Theory of Artificial Intelligence 2021

Toward Out-of-Distribution Generalization Through Inductive Biases

Caterina Moruzzi
Department of Philosophy, Universität Konstanz, 78457, Konstanz, Germany
caterina.moruzzi@uni-konstanz.de

Keywords: inductive biases, generalization, decision making, causality, hybrid AI

Abstract State-of-the-art Machine Learning systems are able to process and analyze a large amount of data but they still struggle to generalize to out-of-distribution scenarios. To use Judea Pearl's words, "Data are profoundly dumb" (Pearl & Mackenzie, 2018); possessing a model of the world, a representation through which to frame reality is a necessary requirement in order to discriminate between relevant and irrelevant information and to deal with unknown scenarios. The aim of this paper is to address the crucial challenge of out-of-distribution generalization in automated systems by developing an understanding of how human agents build models to act in a dynamic environment. The steps needed to reach this goal are described by Pearl through the metaphor of the Ladder of Causation. In this paper, I support the relevance of inductive biases in order for an agent to reach the second rung on the Ladder: that of actively interacting with the environment.

1 Introduction

Diagnosis of skin cancer, autonomous driving, and recommender systems are only a few examples of decision-making processes in which Artificial Intelligence (AI) is increasingly involved. The question that emerges from the fast progress of automated systems in this areas is whether, thanks to the growing availability of data, machines will soon become better than humans at decision-making. While some warn against a future where AI will take over human jobs and personal life, others are skeptical that this can happen, unless machines acquire the ability of organizing the manifold of data they process into models that allow them to deal with uncertainty and generalize to out-of-distribution scenarios.

Machine learning (ML) models are very good at handling within-distribution generalization. Still, observing correlations between data is not enough for successfully dealing with scenarios that lie outside the training distribution. To use Judea Pearl's words, "Data are profoundly dumb" (Pearl & Mackenzie, 2018). Big Data are not enough to be capable of generalization, abstraction, and agency; possessing a robust model of the world, a representation through which to frame reality, is a necessary requirement in order to deal with unknown scenarios. The steps needed to reach this goal are described by Pearl through the Ladder of Causation, a metaphor that Pearl uses to describe the performance of a system (Pearl & Mackenzie, 2018). Pearl argues that current state-of-the-art ML models do not progress beyond the first rung of the Ladder: that of observing the environment and finding statistical correlation between available data.[1] AHow to extract from available knowledge the information needed to abstract to out-of-distribution scenarios is the central challenge that I aim to address in this paper. The key for achieving the necessary generalization skills is a feature that will occupy the central stage: robustness (Bertsimas & Thiele, 2006; Hansen & Sargent, 2011). A central contribution of this proposal is to argue in favour of the relevance of inductive biases in building robustness in decision-making systems. To support this claim, I take inspiration from Daniel Kahneman's dual system of human cognition (Kahneman, 2011) to then extend the discussion to the consideration of how robustness can be implemented in hybrid AI models (Garcez & Lamb, 2020).

I conclude by arguing how strategies for embedding inductive biases into ML systems that are based on statistical learning are not enough to achieve the necessary robustness to generalize to out-of-distribution scenarios. Rather, methods aimed at identifying the causal dependencies between variables, such as graphical causal models, are better suited for this scope.

2 Out-of-distribution Generalization

Statistics is commonly concerned with within-distribution generalization and different strategies are already successfully employed to generalize to unseen data drawn from the same distribution (e.g. logistic regression, stochastic regularization, and Wasserstein DRO). What is still a challenge, is to make ML systems perform out-of-distribution generalization, where the testing distribution is unknown and different from the training set. Various methods to solve this problem have been proposed: stable learning (Kuang et al., 2018), domain generalization (Muandet et al., 2013), and causal learning methods (Peters et al., 2016), among others. Still, despite the advancements in the field, a decisive solution to this problem has not been found, yet.

The problem of abstracting to unknown scenarios is pervasive in ML applications. For state-of-the-art Natural Language Inference models, drawing inferences between pairs that require knowledge about phenomena such as modals, implicatives, conditionals, etc. is still a challenge. For example, as pointed out in a recent work by Kalouli (Kalouli et al., 2020), current deep learning models struggle to recognize how the sentence "The judge believed the tourist arrived" implies the sentence "The judge believed the tourist". Most state-of-the-art Natural Language Processing systems are, indeed, both brittle and spurious, failing when text is added or modified, even if its meaning is preserved (Seo et al., 2016).

Generalizing to unknown scenarios is problematic also in other fields of application of ML. In image classification, for example, classifiers are subject to picking up undesirable correlations during training. To use a frequently cited example, a 2016 work by Ribeiro et al. (Ribeiro et al., 2016) observes how in carrying out the task

[1] In what follows, I will refer to a 'system' as the complex of the actuator of a decision-making process and of the process itself.

of distinguishing between pictures of wolves and huskies the success of a particular classification model was not based on an understanding of the distinction between the two species of animals but, rather, on the different background of the pictures in question. The classifier predicted 'wolf' if there was snow or a light background in the picture and 'husky' otherwise.

The urgency of improving the generalization capabilities of ML systems is even more compelling in applications such as medical diagnosis or autonomous driving, where the mis-identification of a possible cancerous tissue or the mistake of a cyclist for an inert object can lead to serious consequences. It is, therefore, critical to identify the mechanisms through which automated systems of decision-making can become more robust and, accordingly, achieve better abstraction skills.

3 Choice in Decision-making Processes

The aim of the rest of the paper is to get insights into how the out-of-distribution generalization challenge can be addressed. To do so, I start by illustrating a decision-making process in the light of Pearl and Mackenzie's Ladder of Causation. This will allow me to identify some crucial steps that lead agents to achieve higher abstraction and generalization abilities.[2]

The Ladder of Causation is a metaphor used to describe a system's competence. It is structured in three rungs:

Rung 1: Correlation. A system operating on the first rung is a mere observer of what happens in the world. The question the agent asks at this stage is: "What is the probability that y happens, given x?".

Rung 2: Intervention. In order to climb to this level, the agent needs to deliberately interact with the environment and alter it. The question is: "What is the probability that y happens if I *do* x?".

Rung 3: Counterfactuality. Agents that reach this step are able to imagine counterfactual scenarios and to adapt their actions accordingly. The question the agent asks here is: "What is the probability that y' would occur had x' occurred, given that I actually observed x and y?" (Pearl & Mackenzie, 2018).

In this paper, I will focus on the mechanism that enables agents to proceed from Rung 1 to Rung 2, leaving the last step toward the third Rung to other discussions (Moruzzi, 2022). I call this mechanism 'choice' and, in order to make the description of how it works more concrete, I illustrate the process through an example.

Suppose that an agent (I will call her Anne) has a headache and her aim is to stop it. The steps that she will (presumably) take to decide on the best choice in order to achieve this aim follow the three rungs of the Ladder of Causation (see Figure 1).

From the data, or *Knowledge*, that is available to her, the agent begins to build an initial, approximate, frame that helps her understand and organize the data (F_a). In our example, to achieve the aim of getting rid of the headache, Anne may start by observing with which probability the variable 'No headache' is associated to the

[2] In this paper I will not discuss whether artificial systems can be deemed agents and what agency amounts to, leaving this question to further research.

variable (drinking) ‘Coffee’. In other words, she observes with which probability the action of drinking coffee (x) is associated to the effect of stopping the headache (y).

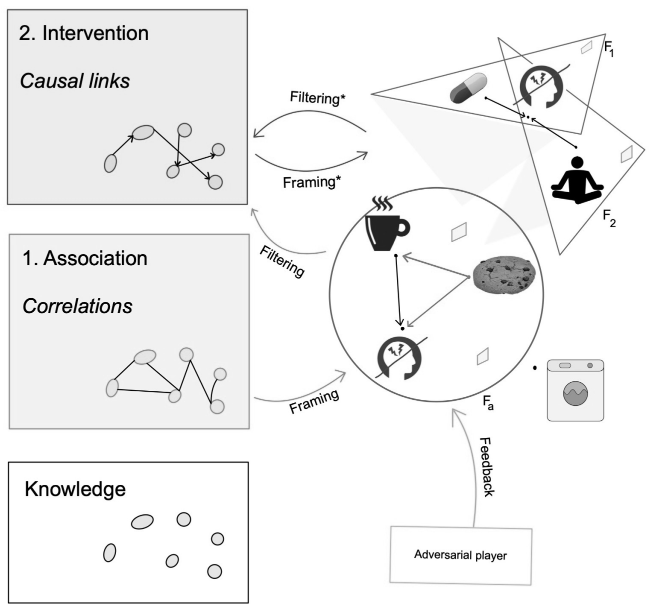

Fig. 1 Decision-making Process

Anne may then proceed at the level of *Intervention* by drinking coffee to confirm the connection between the variables ‘Coffee’ and ‘No headache’, observed in the previous step. Her question is: “What is the probability that the headache will stop if I drink coffee?” If the intervention supports the observed data, the agent can draw causal links between the variables ‘Coffee’ and ‘No headache’.

In order to reinforce the frame that the agent starts to build at the level of Association, she chooses among the alterations to the environment the ones that produce the desired outcome. In our example, Anne will use her frame to *filter* out variables that are not related to the outcome ‘No headache’, for example (doing the) ‘Laundry’. Once the agent observes the frame to work reliably, she starts interpreting data according to it. In turn, the frame influences the agent in drawing causal links between elements in future experiences (*Filtering** in Fig. 1).

The two processes of framing and filtering are tightly interconnected in what I call the mechanism of ‘choice’, as in this process the agent makes the decision of which action to take to maximize her reward. Two are the components that allow the success of the choice mechanism: *feedback* and *inductive biases*. I will address inductive biases in section 4; here I briefly discuss the contribution brought by the feedback component to the mechanism of choice.

The process of optimization and adaptation to other frames is facilitated by the feedback received from an *Adversarial Player*, a notion taken from robust control theory which refers to it also as Malevolent Nature (Hansen & Sargent, 2011). The adversarial play between the agent and the Adversarial Player can be understood as a max-min decision rule, which has the result of improving the robustness of the model (Goodfellow et al., 2014). In our example, suppose Anne needs to stop drinking coffee because her blood pressure is too high. The Adversarial Player may make Anne try different things to get rid of the headache, for example drinking hot milk, taking vitamin pills, practicing relaxation techniques, and so on. Through an active intervention in the world (in Pearl's terms, the do-operator), the agent learns to discriminate between the actions that are successful (relevant information) and those that are not (irrelevant information) and she adapts her frame to the new context. For example, Anne may find out that practicing relaxation techniques works against headache and add it to her frame (F_2). Or, in another context, she may discover that taking an Aspirin works, and add it to yet another frame (F_1). The feedback mechanism and the adaptation to other frames are essential in the process toward building robustness in decision-making and performing counterfactual thinking, a prerequisite to climb toward the third rung of the Ladder of Causation (Moruzzi, 2022).

4 Inductive Biases

In addition to the feedback received from the Adversarial Player, the success of the choice mechanism depends also on inductive biases that the agent already possesses, prior to engaging with the decision-making process in question.

Induction consists in drawing inferences about unknown values of variables on the basis of past observations. The problematic nature of induction is well-known (Vickers, 2014): How can we draw inferences to future occurrences on the basis of past events? Even if the inductive framework we construct is accepted as valid, there can still be contradictory observations that are consistent with our framework. We humans move through the maze of uncertainty through guiding principles that give us some confidence in the fact that our generalizations based on past observations will extrapolate well into unknown ones. These principles are known as inductive biases.

In Anne's example, prior to engaging with the construction of the frame at the level of Association, Anne may already have some priors that are relevant to the selection of the information that will enter her frame. For example, she may start by assuming that the ingestion of some liquids interacts with the body, triggering chemical reactions. This assumption helps her identify 'Coffee' as a potential relevant variable in achieving the aim of stopping the headache.

Assuming that these inductive biases, or assumptions, help agents adapt their frames to previously unseen scenarios, the question is which methods we should use to inject similar inductive biases into automated systems of decision making, in order to address the problem of out-of-distribution generalization. The first step is to consider which role inductive biases play in human decision-making processes. I anticipate that Daniel Kahneman's dual-system theory of human cognition, although controversial, can help making the transition between human and artificial cognition smoother and understanding how the introduction of inductive biases into automated systems can support their climb toward generalization.

In his best-selling book, *Thinking, Fast and Slow* (Kahneman, 2011), Kahneman uses the fictional characters of System 1 and 2 to refer to the two kinds of computation of the brain. System 1 is effortless, fast, and automatic and it is the one that we use the majority of the time to interact with the world. System 2, on the other hand, is slower and more rational. It is activated when we need to perform a task that requires our full attention or when we are confronted with an unexpected situation. I start by trying to understand where these two systems come into play in the example of Anne's decision-making process.

In human cognition, System 2 is responsible for the slow and analytical process of selecting from the data the relevant variables that constitute the agent's initial frame, which allows her to start building a narrative from available knowledge (Kahneman et al., 2021). What guides System 2 in this process are priors, inductive biases that help the agent identify causal dependencies between variables.

Once the frame is built and set, and the agent is at the level of Intervention, System 1 is activated as she needs to react fast in interpreting future scenarios using the available frame. Frames can be considered to be cognitive shortcuts, they allow the agent to focus her mind only on the relevant information, thus reducing cognitive load. The effortlessness of System 1 and its speed in using this frame to interact with the environment help the agent react to changing conditions. The feedback that the agent receives from the environment when engaging in the search of variables that lead to the achievement of her aim, contribute to making the frame more robust.

The fact that the human brain can effortlessly identify causal dependencies in the environment has been interpreted by some as an evolutionary advantage of humans (Tomasello, 2014). However, this advantage may turn into a source of cognitive errors: *harmful biases* may emerge if the agent interprets data on the basis of a wrong model of the world or if she identifies causal relations were there aren't any (Kahneman et al., 1982). Biases can be interpreted as deviations from the correct model due to the fact that an agent applies System 1 thinking to create the links without pausing to interrogate the data.

Going back to Anne's example, suppose that she observes that the probability of the two variables 'Coffee' and (eating) 'Cookies' happening together is high. She may, erroneously, conclude that the variable 'Cookies' is causally related to 'No headache'. This, in Pearl's term, would be a confounder: a variable that is capable of producing spurious associations between the input and outcome, not attributable to their causal dependence (Pearl & Mackenzie, 2018).

In order to identify confounders, the agent necessarily needs to refine her frame. This is made possible by the activation of System 2 during the feedback process. Anne may start doubting that the frame she is using is the correct one. The doubt could originate if, for example, she happened to eat cookies while drinking hot milk and did not get the desired result of stopping the headache. By reflecting on whether the variable ‘Cookies’ should rather stay without of her frame, Anne is employing her analytical skills to delay intuition and to adapt her frame in response to the feedback received from the environment (Kahneman, 2011). This capacity of adapting and refining frames allows the agent to prevent herself from drawing spurious correlations and to make her decision-making process more robust.

5 Hybrid AI Methods

In the previous section, I supported the relevance of inductive biases for the robustness of a decision-making process and for the generalization capabilities of the agent that performs it. Inductive biases are currently introduced into ML systems through various methods, for example dropout (Srivastava et al., 2014) or early stopping, techniques used for addressing overfitting problems, and model selection based on cross-validation.

If inductive biases are understood as a set of rules or assumptions that the agent possesses a priori, in advance of starting the learning process, it seems arguably easier to inject these priors into models that are built following a symbolic approach. For example, systems based on inductive logic programming encode knowledge into inductive biases and induce algorithms to derive hypothesis from a pre-specified database of facts (Muggleton, 1991).

But while symbolic methods can handle extrapolation from pre-encoded information and, as System 2 in human cognition, are better suited to analytically select relevant information from a dataset thanks to inductive priors, they are also brittle and less capable of processing large amounts of data in contrast to sub-symbolic methods. On the other hand, sub-symbolic methods learn faster and, as System 1, can quickly react to changing conditions in the environment. However, they are prone to output spurious results and to let their decision-making process be guided by undesirable correlations between variables in the training set (as noted in section 2).

Some claim that a way to solve this problem is to support the data-processing skills of ML models through the capacity of abstraction and logical reasoning of symbolic AI methods. This solution is the one proposed by the hybrid approach in AI research, which is showing promising results in many areas (Bahdanau et al., 2018; Bengio et al., 2019; Booch et al., 2020; Garcez & Lamb, 2020; Madan, Ke, Goyal, Schölkopf, & Bengio, 2021; Moruzzi, 2020). This hybrid approach of symbolic and sub-symbolic methods would arguably allow to hold the advantages of both strategies, get rid of their respective weaknesses and, at the same time, program models that fare much better in generalization and abstraction.

Many of the conceptualizations of state-of-the-art hybrid AI methods have been inspired by Kahneman’s dual-system theory of the human mind (Kahneman, 2011). As mentioned above, the way System 1 works can be compared to the mechanisms of sub-symbolic AI, which is essentially learning from experience and can react fast. On the other hand, System 2’s capability of solving complex problems can be associated with techniques based on logic and planning that employ explicit knowledge and reasoning with symbols.

The strength of state-of-the-art hybrid AI models is their capacity of combining the computational power of Deep Learning with symbolic and logical reasoning to not only process large amounts of data, but also to identify which elements within those data stay stable (Bengio, 2017). To be able to handle dynamic, changing conditions, without letting decision-making processes be subject to spurious correlations and harmful biases, a possible way forward could, thus, be found in the combination of statistical models, able to perform System 1 tasks and *fast* in adapting the initial frames to the changing conditions of the environment, with symbolic models, able to perform System 2 tasks and construct frames that help the agent interpret the environment and interact with it.

Strategies to build inductive biases into ML systems are already widely adopted (i.e. linear models, decision trees, naive Bayes). However, while they work well for within-distribution generalization, they struggle to generalize to interventions that lie outside of the training distribution. One of the reasons for this weakness is that these methods are usually based on the observation of correlation among data, without considering causal dependencies between variables. In Pearl's terms, they are stuck at Rung 1. A strategy that could help automated systems organize the information into frames and climb up the ladder toward Rung 2 are graphical causal models, graphs which represent probabilistic dependencies between variables and which can be used by agents to organize their beliefs regarding causal structures (Pearl, 2000). The structure of causal graphs remains invariant, even if the situation changes. The ensuing advantage is that, by identifying the variables that are responsible for change and that remain stable through varying conditions, the agent can more easily adapt her frames to new and out-of-distribution situations (Pearl & Mackenzie, 2018).

Indeed, as suggested in (Eva et al., 2019), identifying similarities between causal graphs can help the system adapt its behaviour to unknown scenarios and to predict the outcomes of her actions. To use an overly simplified example, if Anne identifies what is that the causal graph that connects 'Coffee' to 'No headache' has in common with the graph that connects 'Aspirin' to 'No headache', she may be able to extrapolate the information she needs in order to identify other, unknown variables that can stop her headache. By identifying the dependencies between variables that remain stable in dynamic context, the agent can, at the same time, generalize to out-of-distribution interventions and climb toward the highest rungs of the Ladder of Causation (Moruzzi, 2022).[3]

6 Conclusion

While the introduction of inductive biases into ML models is already adopted by practitioners and has been addressed in past research (see section 4), strategies employed so far do not perform well in generalization tasks. I suggested that graphical causal models could represent a viable strategy to enable the move toward higher

[3] I thank the reviewers for advising me to develop more this section.

rungs of the Ladder of Causation, by allowing systems to identify structures of dependencies between variables which can be used to forecast out-of-distribution scenarios. The contribution that this paper brings to debates on generalization in ML is twofold. First, the illustration of the decision-making process in the light of Pearl's Ladder of Causation allows to clearly identify the steps required in order to proceed from the mere observation of variables to the interaction with the environment and its modification. Secondly, it identifies the discrimination between harmful biases and beneficial assumptions, namely inductive biases, a key component toward making the decision-making process more robust, by drawing on the research conducted by Kahneman's research on errors and biases in human cognition (Gilovich et al., 2002; Kahneman, 1973). Studying the role of inductive biases in supporting human and automated systems to frame information can shed light on promising strategies which can contribute to the development of artificial decision-making systems that perform better in out-of-distribution generalization (Bender & Friedman, 2018).

References

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., & Courville, A. (2018). Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. doi: 10.1162/tacl_a_00041

Bengio, Y. (2017). The consciousness prior. *arXiv preprint arXiv:1709.08568*.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., ... Pal, C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

Bertsimas, D., & Thiele, A. (2006). Robust and data-driven optimization: modern decision making under uncertainty. In *Models, methods, and applications for innovative decision making* (pp. 95–122). INFORMS.

Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., ... others (2020). Thinking fast and slow in ai. *arXiv preprint arXiv:2010.06002*.

Eva, B., Stern, R., & Hartmann, S. (2019). The similarity of causal structure. *Philosophy of Science*, *86*(5), 821–835.

Garcez, A. d., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd wave. , *arXiv eprint arXiv:2012.05876*.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hansen, L. P., & Sargent, T. J. (2011). *Robustness*. Princeton, NJ: Princeton University Press.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). N.J: Prentice-Hall: Englewood Cliffs.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Macmillan.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. New York: Little, Brown & Co.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kalouli, A.-L., Crouch, R., & de Paiva, V. (2020, December). Hy-NLI: a hybrid system for natural language inference. In *Proceedings of the 28th international conference on computational linguistics* (pp. 5235–5249). doi: 10.18653/v1/2020.coling-main.459

Kuang, K., Cui, P., Athey, S., Xiong, R., & Li, B. (2018). Stable prediction across unknown environments. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1617–1626).

Madan, K., Ke, N. R., Goyal, A., Schölkopf, B., & Bengio, Y. (2021). Fast and slow learning of recurrent independent mechanisms. , *arXiv preprint arXiv:2105.08710*.

Moruzzi, C. (2020). Artificial creativity and general intelligence. *Journal of Science and Technology of the Arts*, *12*(3), 84–99.

Moruzzi, C. (2022). Climbing the ladder: How agents reach counterfactual thinking. In *ICAART 2022: 14th international conference on agents and artificial intelligence* (pp. 555–560).

Muandet, K., Balduzzi, D., & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International conference on machine learning* (pp. 10–18).

Muggleton, S. (1991). Inductive logic programming. *New generation computing*, *8*(4), 295–318.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Hachette UK.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(5), 947–1012.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Tomasello, M. (2014). *A natural history of human thinking*. Cambridge, MA: Harvard University Press.

Vickers, J. (2014). The problem of induction. *Stanford Encyclopedia of Philosophy*.