November 2022

Perspective
EXPERT INSIGHTS ON A TIMELY POLICY ISSUE

EVAN D. PEET, BRIAN G. VEGETABILE, MATTHEW CEFALU, JOSEPH D. PANE, CHERYL L. DAMBERG

Machine Learning in Public Policy

The Perils and the Promise of Interpretability

Major advances in statistics and computer science in recent years have yielded increasingly advanced tools in artificial intelligence (AI) and the subdiscipline of machine learning (ML). ML algorithms are "methods that can automatically detect patterns in data, and then . . . use the uncovered patterns to predict future data or other outcomes of interest" (Murphy, 2012). The advances in ML coincide with the ever-growing amount of "big data" collected by government agencies and private companies or voluntarily submitted by individuals. ML tools are particularly well suited to leverage these big data, automate analyses, and produce results that inform critical decisions.

For these reasons, numerous entities are making enormous investments in ML and working to rapidly apply ML for a variety of uses. One area of high-impact use is in informing public policymaking and the design of programs. ML can have a significant impact on public policy by modeling complex relationships, improving policy design, augmenting human decisionmaking, and enhancing the speed and quality of public services (Berryhill et al., 2019). However, ML has the potential for

RAND CORPORATION

misuse due to overconfidence in its promise and a potential lack of interpretability. The misuse of ML tools introduces new dimensions of risk, particularly if the algorithms are not properly understood.

In this Perspective, we provide an overview of ML interpretability and how the associated methods and tools can be used to harness the promise and avoid the potential perils of ML. We start by defining ML interpretability, then describe methods that aid interpretation and the characteristics of effective explanations. Finally, we conclude with recommendations for policymakers.

Promise of Machine Learning in Public Policy

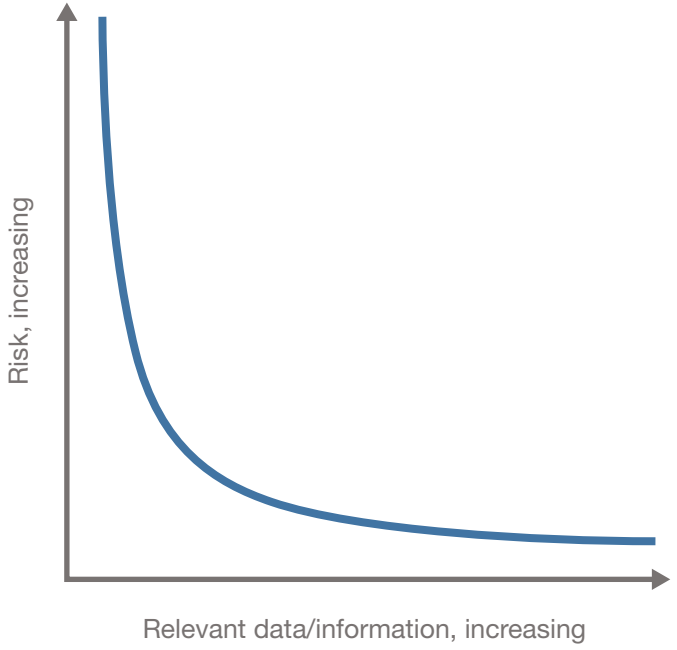
Already, ML is leading to advances in labor productivity (Damioli, Van Roy, and Vertesy, 2021), human safety (e.g., autonomous vehicles) (Wang et al., 2020), and other aspects of society. Included among these other aspects of society is public policy. The use of ML can play out in small, local policy decisions as well as at a national policy scale. For instance, some local governments have begun to employ ML to examine social media posts regarding restaurant cleanliness to target health department inspections (Kang et al., 2013). At the national level, governments have begun using ML to assign refugees to locations based on recent research on refugee integration (Bansak et al., 2018). ML has also been used to efficiently aggregate and process various streams of data to support disaster response (Khatoon et al., 2022), to predict recidivism risk and inform bail decisions in criminal justice (Dressel and Farid, 2018; Elyounes, 2020), to predict patient risk of readmission and other health outcomes (Ahmad, Eckert, and Teredesai, 2018), to inform early identification of at-risk students (Hu and Rangwala, 2020), and to inform screening and decisionmaking in cases of child maltreatment and other social services (Chouldechova et al., 2018). Recent advances in ML demonstrate the ability of these algorithms to learn quickly and to handle complex, strategic processing (Bory, 2019). Many more applications of these methods, including public policy applications, will be explored in the future.

Policymakers use scientific evidence to inform their policy decisions, but until recently, the evidence typically used by policymakers has been based on small, lagged (or old) data and relatively slow and/or simple analytic methods (e.g., linear models). However, the actual relationships between the many variables of interest to policymakers are often complex (e.g., the many factors and complex interactions that lead to climate change) and might not be accurately reflected in standard, simple models (Mitchell, 2009). This is why ML algorithms often lead to better-fitting (or more-accurate) models and predictions (Bell et al., 2022).

ML has the potential to advance public policy analysis in multiple ways. First, as depicted in Figure 1, by leveraging the increasing amount of available information, ML has the potential to more accurately predict policy outcomes, thereby reducing the risks of unintended consequences (Rasouli and Timmermans, 2014). Also, with outcome predictions, ML can advance the analysis of public policy problems by uncovering heterogeneity and equitably allocating scarce resources to those at greatest risk of the worst outcomes (Schultz, Lovejoy, and Peet, 2019). Furthermore, ML algorithms can automate data processes to offer more-rapid predictions that increase the possibility of early intervention (Ruiz et al., 2019). Finally,

FIGURE 1

Hypothetical Relationship Between Risk and Information



ML can also advance public policy analysis by leveraging new data from unconventional sources (e.g., social media, images, and audio) and use nonlinear modeling to uncover previously unknown and complex relationships (Kino et al., 2021). A more accurate understanding of complex relationships among the varied stakeholders involved in public policy has the potential to reveal the implications of policies, thus informing policymakers' aims to achieve better, principled, objective, and equitable policy decisions (Coyle and Weller, 2020).

Peril of Machine Learning in Public Policy

The potential promise of ML also comes with trade-offs that, if not handled properly, can lead to peril. For instance, overconfidence in how the results of an ML algorithm would perform in different contexts than those observed in the data can lead to costly investment losses (Zillow, 2021). More perilously, incorrectly applied and interpreted ML algorithms can perpetuate structural racial and ethnic inequities, as has occurred in health care when ML algorithms led to healthier White patients receiving additional care at the expense of sicker Black patients (Obermeyer et al., 2019). Similar racial and ethnic and socioeconomic inequities resulted from the UK Education Department's attempt to rescale the inflated grades that occurred during the pandemic to those of previous years. In this case, high-achieving students in low-income schools were downgraded to match their schools' expected outcomes, while students from affluent private schools kept their inflated grades because their schools' sizes were too small to calculate alternative scales (Coyle, 2020). Additionally, when using ML to recruit new hires, the same biases against women that exist in the labor force and in labor data can be perpetuated with inappropriate modeling (such as defining a successful hire with potentially biased measures like salary) (Florentine, 2018). These and other instances of ML failures involved overconfidence in the results and the trade-off between accuracy and interpretability.

Overconfidence might arise from the novelty of ML, and the belief that these new methods can magically produce high-quality results without the biases or contextual limitations present in the data. As with any sta-

tistical method, the results are only as good as the inputs. Although ML methods have demonstrated the promising ability to rapidly process complex data that is necessary in policy settings (Bory, 2019), all results reflect the context and quality of inputs. For instance, in closed systems with defined boundaries and objectives, such as in the board game Go, ML has demonstrated promise (DeepMind, 2020). In contrast, when new and different contexts arise, users might be overconfident in ML results. For instance, the rapid and unprecedented opioid crisis means that the historical data used with ML to predict risk of opioid misuse likely underestimates actual risk (Kilby, 2021). Public policies are similarly implemented in complex and continually evolving contexts that might not have historical precedent observed in data.

Another potential peril involves the trade-off between accuracy and *interpretability* by those who will act on the information generated from ML models. ML algorithms offer the promise of more-accurate predictions by modeling outcomes as an unspecified function of variables. In cases in which theory or empirical evidence does not specify how a variable could affect an outcome, ML algorithms can automatically detect whether the variable is important and the nature of its relationship with the outcome. This ability to focus on the important variables and capture complex relationships among variables without an *a priori* definition offers more flexibility than standard (i.e., linear) methods and can lead to more-accurate predictions. Even standard statistical methods like linear models with a small number of variables can be difficult for humans to interpret without statistical knowledge or training. ML algorithms magnify the interpretation difficulties with nonlinearities and interactions among many different variables. The complexity of the ML models and their outputs create difficulties with interpretation of the results and have led ML to gain the moniker *black box*—a system whose internal workings are obscured. This moniker is accompanied by the perception that ML lacks transparency, accountability, and trustworthiness. A black box without transparency, accountability, and trustworthiness is particularly problematic in the policy arena, as policymakers must gain the confidence of the public, making difficult decisions and providing clear, easy-to-understand rationales despite the fact that policies might not benefit all parties.

Harnessing the Promise, Avoiding the Peril

Although ML algorithms "automatically detect patterns in data" to enable the analysis of big data and advance areas of research where standard methods have struggled to understand complex relationships among many variables and accurately predict outcomes, as with any tool, ML should be applied only when appropriate (Murphy, 2012). Inappropriately used, black-box ML algorithms could fail in policy settings by not providing the information needed (such as for whom, for what, and when the results apply) for robust, risk-aware decisions.

Interpretability

Information that is trustworthy, transparent, and accountable is needed to make fair, equitable, and risk-aware policy decisions that have significant implications for those affected. The goal of interpretable ML is to provide trust-

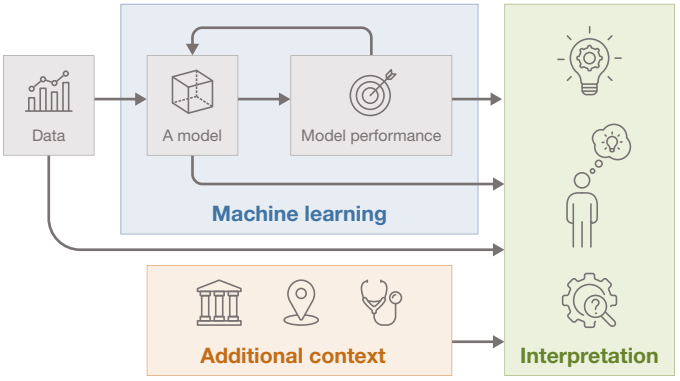
worthy, transparent, and accountable information that leads to greater understanding of the problem and implications of different solutions. However, understanding is personal and might have different meanings for each audience (Preece et al., 2018). This leads to an air of "I'll know it when I see it" around interpretability (Lipton, 2016).

As illustrated in Figure 2, interpretability requires multiple types of information. The first type of information required for interpretability is the data. Many of the previously mentioned ML failures could have been avoided with a better understanding of the data, such as an acknowledgment that the data generated from electronic health records reflect structural inequities in access.

The next type of information required for interpretability concerns the model. The best models of how an outcome is related to other variables are accurate simplifications of reality. However, models also involve assumptions and could be biased if the assumptions are wrong. Standard linear models have straightforward interpretations (e.g., for each unit increase in variable *X*, the outcome *Y* changes by *Z* units, if all other factors were held constant). In contrast, ML models capture nonlinear relationships between variables (e.g., if all other factors were held constant, for each unit increase in variable *X* in the range of values *A* to *B*, the outcome *Y* changes by *Z* units, and for each unit increase in *X* in the range of values *C* to *D*, the outcome *Y* changes by *W* units). Cases of diminishing returns are examples of this, as the impact of *X* on *Y* diminishes at higher values of *X*. The interpretations are further complicated by the interactions captured by ML models (e.g., if all other factors were held constant, for each unit increase in variable *X*, the outcome *Y* changes by *Z* units if *S* is present; if *S* is not present, each unit increase in *X* changes the outcome *Y* changes by *V* units). ML models can also capture nonlinear interactions, further complicating the interpretation.

FIGURE 2

Types of Information Required for Interoperability



Other important types of information that are required for ML interpretability are model performance and additional context. Model performance metrics describe the accuracy of the model's predictions. Less accuracy leads to less trust in the model. Additional context about relevant factors (e.g., geography) and limitations (e.g., sample representation) might also be necessary for interpretability and provide guidance for how the results might apply to other questions.

Information about the data, the model and model performance, and the additional context are each critically important for interpreting the results of any analysis. However, while other more standard types of analyses are familiar to audiences, the novelty and complexity of ML algorithms can make their results more difficult to interpret.

Model Interpretability Methods

To aid the interpretation of ML models, researchers have been developing interpretability methods (Arrieta et al., 2020). These methodological innovations can be used to convey information so that multiple audiences, including policymakers, can quickly and easily understand the relationships between variables. These methods can be divided into two types: (1) those that facilitate intrinsic interpretability and (2) those that facilitate post hoc interpretability.

Intrinsic Interpretability

Intrinsic interpretability means that, by design, the ML model is constructed to enable a level of interpretability for certain audiences. For example, some audiences perceive linear models to be intrinsically interpretable because their training has taught them the associated statistical theory and how to understand the model coefficients. However, ML models are also intrinsically interpretable if they are (1) simulatable (i.e., a human can contemplate the entire model at once), (2) decomposable (i.e., each component of the model has an intuitive explanation or decomposes into understandable components), and (3) transparent (i.e., there is clear documentation with known mathematical properties) (Lipton, 2016).

As an example of intrinsically interpretable ML, consider the optimal rule list models developed by Angelino and colleagues, 2017, and Lakkaraju, Bach, and Leskovec, 2016. Rule lists are a set of *if-then* statements that divide the input variables into decision trees (e.g., if location = Eastern United States, then predict average temperature = 65 degrees F). These rule list models are simulatable because they focus on the most-easily conceptualized input variables; they are decomposable because they are composed of intuitive *if-then* statements; and they are transparent because the algorithm operates by minimizing the difference between the model's prediction and the observed data. Another example of intrinsically interpretable ML is the deep learning image recognition models developed by Chen and colleagues, 2019. This ML model uses *prototypical examples* from the training data (e.g., if an object has a beak like the provided examples, then it is a bird) to classify images. This algorithm is simulatable and decomposable because of its focus on understanding the parts of the image that are contributing to the outcome, and it is transparent because it is optimized using a well-defined approach.

Post Hoc Interpretability

For nonintrinsically interpretable ML models, an understanding of the model can be achieved using post hoc interpretability methods. These methods are not incorporated in the model; rather, they require an additional stage of analysis. Although we describe a few examples here, there are many other post hoc interpretability methods (e.g., local interpretable model-agnostic explanations and accumulated local effects). Each of the methods attempts to wrestle with the results of the ML model, determining how they might be applied in certain contexts and what their limitations are.

Take, for example, the popular random forest ML model. *Random forests* are an ensemble ML model (i.e., many models combine to predict an outcome) that leverages randomization to create a diverse "forest" of decision tree models by training a large number of individual tree models, each on random subsets of the data and each using

only a random subset of features to learn relationships. The power of this method is that the diversity obtained through randomization provides an ability to model complicated outcomes when averaging across the many trees in a forest. Random forests are not intuitive and can be very complex. The results are not intrinsically interpretable, but post hoc methods that can summarize the results have been developed. For instance, a common post hoc interpretability method for random forest models describes the *variable importance*, or the relative contribution of each variable to the prediction. Another popular post hoc interpretability method is the partial dependence plot. In the same way the coefficients from a linear regression describe how changes in X affect Y, partial dependence plots describe these relationships but with more complexity. Instead of having one number summarizing how much Y changes with a one-unit change in X, partial dependence plots visualize the relationship and allow for nonlinearities (e.g., demand for winter coats, which is steady for a range of low temperatures until eventually declining as temperature increases). (See Molnar, 2020, for more details on partial dependence plots and other post hoc interpretability methods.)

Effective Explanations

Explanations should, as shown in Figure 2, describe the data, model, model performance, and additional context. In describing the model, both intrinsic and post hoc interpretability methods can be used. However, what makes an explanation of ML effective is not the specific interpretability methods used. Rather it is an understanding of the audience and how the information can be adapted to meet the needs of the audience (Preece et al., 2018).

What makes an explanation of ML effective is an understanding of how the information can be adapted to meet the needs of the audience.

Evidence from large bodies of research in philosophy, psychology, and cognitive science describe four characteristics of effective explanations based on how humans communicate, generate, select, evaluate, and ultimately trust the information they receive. The same characteristics of trusted human-human communication can be applied to explanations of ML models. In this Perspective, we describe effective explanations as (1) *conversational*, (2) contrastive, (3) *selective*, and (4) *pragmatic* (Miller, 2019).

First, trusted human-human communication is often *conversational*. Conversations are a form of communication that can assume some level of misunderstanding (if everything is already known or understood, why ask?) and allow for follow-up questions and clarifications (Miller, 2019). Conversations are iterative, including back-and-forth dialogue between participants. Conversations allow for each participant to understand each other better, thus enabling the adaptation of the message to meet the needs of the audience. Conversational explanations regarding

ML models might involve multiple modes (e.g., written reports and interactive presentations) that are iterative and allow the audience to gradually build their understanding, receiving information and responding with probes exploring the model's limitations (Morrison et al., 2018).

Additionally, one of the most important findings in the philosophical and cognitive science literature is that people typically ask *contrastive* questions, exploring why something happened instead of something else. Contrastive explanations limit the scope of the explanation to understanding the difference among a small number of cases rather than describing all the causes of the event and their relative importance (Lipton, 1990). For example, a contrastive question could be, "Why does the model predict that David, not Mary, will readmit to the hospital within ten days?" Post hoc interpretability methods, such as partial dependent plots, can be helpful in addressing contrastive questions (e.g., "David is more likely than Mary to readmit to the hospital within ten days because he has a diagnosed heart condition while she does not").

Although interpretability involves transparency, too much transparency can overwhelm, leading to the results being ignored.

Although interpretability involves transparency, too much transparency can overwhelm, leading to the model and the results being ignored. For this reason, trusted human communications are also *selective*. Selection involves the ability to discuss the complexities of the methods in layman's terms, essentially teaching the basics to those whose expertise lies elsewhere and finding the right balance of how much information to share (Morrison et al., 2018). While difficult, is imperative for researchers to understand and be able to provide intuitive descriptions of complex ML algorithms. The specific information to share will vary by audience (e.g., focusing on variables that can be affected by policy rather than others that might be more predictive but cannot be changed) and be informed by the audience's interaction (another reason for conversational communication).

People also tend to trust *pragmatic* explanations. Explanations that are simple, more broadly applicable (Lombrozo, 2007), and consistent with prior knowledge (Thagard, 1989) are more readily accepted than their alternatives. Unfortunately, these preferences do not always lend themselves to truth. Although simple is preferred, the truth might be nuanced, complicated, and difficult to communicate. One way to communicate difficult concepts is through visualization, which has the ability to translate abstract concepts—like probabilities—and make them concrete (Berinato, 2016). Several ML interpretability methods, such as partial dependence plots, offer the opportunity to visually explain abstract concepts.

Recommendations for Public Policy

ML's black-box reputation is not unwarranted; previous applications of ML have failed (Zillow, 2021) and perpetuated inequities in health care (Obermeyer et al., 2019), education (Coyle, 2020), and the labor market (Florentine, 2018). These failures came about because of overconfidence in the promise of ML and a lack of interpretability. However, these failures should not preclude the use of and improvement in ML for public policy. Careful application of ML as a tool for understanding policy problems and testing solutions offers the promise of better, more-objective, and more-equitable policy decisions (Coyle and Weller, 2020).

Although ML, like all statistical methods, has shortcomings and should not be used inappropriately, the black-box potential of the methods implies a greater need for interpretability. The following three recommendations describe how policymakers can improve, approach, and leverage interpretable ML to harness its promise and avoid its potential perils.

- **Improve data through coordinated investments.** Understanding the underlying data that will feed the model, including the reflection of historical inequities and/or contextual limitations contained within, is a key element of interpretability. And because ML is not magic, the results are only ever as good as the data inputs. Some data limitations can be addressed with more data (e.g., linking electronic health records to social service and other records can provide information about social determinants of health to health care providers [Schultz, Lovejoy, and Peet, 2019]). Other data limitations must be addressed with forethought and a dramatic revision of existing data collection efforts to improve quality (Center for Antiracist Research, 2022). Improving both the quantity and quality (i.e., completeness and lack of bias) of data requires coordinated investments to enable data from different sources to be linked and fill in gaps. However, the return on these investments could include the uncovering of previously unknown and complex relationships (Kino et al., 2021), as well as more-rapid and effective interventions (Ruiz et al., 2019).
- **Approach ML expecting interpretability, and be critical.** Policymakers must approach results produced by ML expecting interpretability—that is, expecting that the researchers can provide information on the data, the model and its performance, and the additional context (e.g., limitations) that policymakers need to make robust, risk-aware decisions. This means that policymaker audiences should critically assess data quality, modeling assumptions, and other contextual factors that lead to the results. In doing so, policymaker audiences should demand communication with researchers that provides transparent, trustworthy, and accountable information using conversational, contrastive, selective, and pragmatic communication (Miller, 2019).
- **Leverage interpretable ML to understand policy values and predict policy impacts.** With the promise of more-accurate predictions, interpretable ML could enable policymakers to better understand policy impacts prior to implementation (Rasouli and Timmermans, 2014). Better policy could also

be achieved by a more-informed allocation of scarce resources to those to those at greatest risk of the worst outcomes (Schultz, Lovejoy, and Peet, 2019). In addition, public policy could be improved by providing more insights into the implicit weights received by the various objectives being addressed by each policy and how the weights of each objective could be revised to improve fairness and equity (Coyle and Weller, 2020).

Achieving the promise of ML in public policy depends on the interpretability of each element, from data to end results. Although there have been failures, ML has also produced successes in the areas of refugee integration (Bansak et al., 2018), disaster response (Khatoon et al., 2022), criminal justice (Dressel and Farid, 2018), health care (Ahmad, Eckert, and Teredesai, 2018), education (Hu and Rangwala, 2020), and social services (Chouldechova et al., 2018). With interpretability, ML can achieve its promise of leading to fairer and more-equitable public policy decisionmaking.

References

Ahmad, M. A., C. Eckert, and A. Teredesai, "Interpretable Machine Learning in Healthcare," *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, September 2018.

Angelino, E., N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, "Learning Certifiably Optimal Rule Lists for Categorical Data," *Journal of Machine Learning Research*, Vol. 18, No. 234, 2017.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," Information Fusion, Vol. 58, June 2020.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, and J. Weinstein, "Improving Refugee Integration Through Data-Driven Algorithmic Assignment," *Science*, Vol. 359, No. 6373, 2018.

Bell, A., I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-Off in Machine Learning for Public Policy," *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 2022.

Berinato, S., *Good Charts: The HBR Guide to Making Smarter, More Persuasive Data Visualizations*, Harvard Business Review Press, 2016.

Berryhill, J., K. K. Heang, R. Clogher, and K. McBride, "Hello, World: Artificial Intelligence and Its Use in the Public Sector," OECD Working Papers on Public Governance No. 36, November 2019.

Bory, P., "Deep New: The Shifting Narratives of Artificial Intelligence from Deep Blue to AlphaGo," *Convergence*, Vol. 25, No. 4, 2019.

Center for Antiracist Research, *Toward Evidence-Based Antiracist Policymaking: Problems and Proposals for Better Racial Data Collection and Reporting*, Boston University, June 2022.

Chen, C., O. Li, C. Tao, A. Jade Barnett, J. Su, and C. Rudin, "*This* Looks Like *That*: Deep Learning for Interpretable Image Recognition," *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.

Chouldechova, A., D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions," *Proceedings of the First Conference on Fairness, Accountability and Transparency*, 2018.

Coyle, D., "The Tensions Between Explainable AI and Good Public Policy," Brookings TechStream, September 15, 2020.

Coyle, D., and A. Weller, "'Explaining' Machine Learning Reveals Policy Challenges," *Science*, Vol. 368, No. 6498, 2020.

Damioli, G., V. Van Roy, and D. Vertesy, "The Impact of Artificial Intelligence on Labor Productivity," *Eurasian Business Review*, Vol. 11, No. 1, 2021.

DeepMind, "AlphaGo," website, 2020. As of September 16, 2020: https://deepmind.com/research/case-studies/alphago-the-story-so-far#our_approach

Dressel, J., and H. Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances*, Vol. 4, No. 1, 2018.

Elyounes, D. A., "Bail or Jail? Judicial Versus Algorithmic Decision-Making in the Pretrial System," *Columbia Science and Technology Law Review*, Vol. 21, No. 2, 2020.

Florentine, S., "Amazon's Biased AI Recruiting Tool Gets Scrapped," *CIO*, October 19, 2018.

Hu, Q., and H. Rangwala, "Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students," *Proceedings of the 13th International Conference on Educational Data Mining*, 2020.

Kang, J. S., P. Kuznetsova, M. Luca, and Y. Choi, "Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, October 2013.

Khatoon, S., A. Asif, M. M. Hasan, and M. Alshamari, "Social Media-Based Intelligence for Disaster Response and Management in Smart Cities," in P. M. Pardalos, S. T. Rassia, and A. Tsokas, eds., *Artificial Intelligence, Machine Learning, and Optimization Tools for Smart Cities*, Springer, 2022.

Kilby, A. E., "Algorithmic Fairness in Predicting Opioid Use Disorder Using Machine Learning," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Kino, S., Y.-T. Hsu, K. Shiba, Y.-S. Chien, C. Mita, I. Kawachi, and A. Daoud, "A Scoping Review on the Use of Machine Learning in Research on Social Determinants of Health: Trends and Research Prospects," *SSM-Population Health*, Vol. 15, September 2021.

Lakkaraju, H., S. H. Bach, and J. Leskovec, "Interpretable Decision Sets: A Joint Framework for Description and Prediction," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Lipton, P., "Contrastive Explanation," Royal Institute of Philosophy Supplement, *Philosophy*, Vol. 27, 1990.

Lipton, Z. C., "The Mythos of Model Interpretability," *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016.

Lombrozo, T., "Simplicity and Probability in Causal Explanation," *Cognitive Psychology*, Vol. 55, No. 3, 2007.

Miller, T., "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, Vol. 267, February 2019.

Mitchell, S. D., *Unsimple Truths: Science, Complexity, and Policy*, University of Chicago Press, 2009.

Molnar, C., *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, independently published, 2022.

Morrison, C., K. Huckvale, B. Corish, R. Banks, M. Grayson, J. Dorn, A. Sellen, and S. Lindley, "Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis: Reflections on Communicating Machine Learning in Practice," *ACM Transactions on Interactive Intelligent Systems*, Vol. 8, No. 2, 2018.

Murphy, K. P., *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, Vol. 366, No. 6464, 2019.

Preece, A., D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in Explainable AI," arXiv preprint arXiv:1810.00184, 2018. As of September 17, 2020: https://arxiv.org/abs/1810.00184

Rasouli, S., and H. J. Timmermans, "Using Ensembles of Decision Trees to Predict Transport Mode Choice Decisions: Effects on Predictive Success and Uncertainty Estimates," *European Journal of Transport and Infrastructure Research*, Vol. 14, No. 4, 2014.

Ruiz, V. M., L. Saenz, A. Lopez-Magallon, A. Shields, H. A. Ogoe, S. Suresh, R. Munoz, and F. R. Tsui, "Early Prediction of Critical Events for Infants with Single-Ventricle Physiology in Critical Care Using Routinely Collected Data," *Journal of Thoracic and Cardiovascular Surgery*, Vol. 158, No. 1, 2019.

Schultz, D., S. L. Lovejoy, and E. D. Peet, *Examining Interventions to Address Infant Mortality in Allegheny County, Pennsylvania*, RAND Corporation, RR-A858-1, 2019. https://www.rand.org/pubs/research_reports/RRA858-1.html

Thagard, P., "Explanatory Coherence," *Behavioral and Brain Sciences*, Vol. 12, No. 3, 2989.

Wang, J., L. Zhang, Y. Huang, and J. Zhao, "Safety of Autonomous Vehicles," *Journal of Advanced Transportation*, Vol. 2020, October 2020.

Zillow, "Zillow 2021 Q3: Shareholder Letter," Zillow Group, November 2, 2021.

Acknowledgments

The authors thank Joshua Snoke, Jeanne Ringel, and Paul Koegel for their helpful comments and input on this Perspective.

About This Perspective

This Perspective provides an overview of machine learning in informing public policymaking and the design of programs. We describe the promise of machine learning, including modeling complex relationships, improving program design, augmenting human decisionmaking, and enhancing the speed and quality of public services. We also discuss machine learning's potential perils, specifically involving interpretability. We define machine learning interpretability, then describe methods that aid interpretation and the characteristics of effective explanations. We conclude with recommendations for how policymakers can improve, approach, and leverage interpretable machine learning to harness its promise and avoid potential perils.

This work was funded by gifts from RAND supporters and income from operations and carried out within the Access and Delivery program of RAND Health Care.

RAND Health Care

RAND Health Care, a division of the RAND Corporation, promotes healthier societies by improving health care systems in the United States and other countries. We do this by providing health care decisionmakers, practitioners, and consumers with actionable, rigorous, objective evidence to support their most complex decisions. For more information, see www.rand.org/health-care, or contact

RAND Health Care Communications
1776 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
(310) 393-0411, ext. 7775
RAND_Health-Care@rand.org

About the Authors

Evan D. Peet is an economist at the RAND Corporation, professor at the Pardee RAND Graduate School, and co-director of the RAND Center for Causal Inference. An expert in advanced statistical methods such as machine learning, his research focuses on health policy.

Brian G. Vegetabile is a statistician formerly of the RAND Corporation and currently a data scientist at LinkedIn.

Matthew Cefalu is a statistician formerly of the RAND Corporation and currently a director of data science at Disney Streaming.

Joseph Pane is a statistical analyst formerly of the RAND Corporation and currently a data scientist at Thermo Fisher Scientific.

Cheryl L. Damberg is a principal senior researcher and Distinguished Chair in Health Care Payment Policy at the RAND Corporation. She is also a professor at the Pardee RAND Graduate School. She is a nationally renowned expert in health economics and public policy.

www.rand.org

PE-A828-1