UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

CLTC WHITE PAPER SERIES

Improving the Explainability of Artificial Intelligence

THE PROMISES AND LIMITATIONS OF COUNTERFACTUAL EXPLANATIONS

ALEXANDER ASEMOTA

CLTC WHITE PAPER SERIES

Improving the Explainability of Artificial Intelligence

THE PROMISES AND LIMITATIONS OF COUNTERFACTUAL EXPLANATIONS

ALEXANDER ASEMOTA

June 2024

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Contents

Executive Summary

As AI rises in prominence across domains, it is crucial that companies, governments, and the public understand how AI is impacting decision-making. However, there continues to be a dearth of guidance on how decisions made by algorithmic systems should be explained to those affected by them. This paper seeks to offer some perspective on *counterfactual explanations*, a methodology that has the potential to greatly improve access to recourse for algorithmic subjects.

Counterfactual explanations (CFEs) are intuitive for users because they describe how changing the factors that went into an algorithm-based decision would lead to a different output. Consider the example of an applicant denied admission to a university: a counterfactual explanation might recommend that the applicant increase their test scores or take additional courses to improve their chances. Counterfactual explanations are promising as they provide specific recommendations to the user in a format that does not require significant knowledge of AI.

Although CFEs offer some improvements over other explanation methods, they still have significant limitations. Companies, lawmakers, researchers, and regulators must keep these limitations in mind when considering how and when to use CFEs.

RECOMMENDATIONS FOR REGULATORS

- Avoid requiring AI developers to provide counterfactual explanations as they exist currently, due to deficiencies in existing methodologies.
- Support the development of frameworks for explainability based on domain knowledge and department needs, rather than relying on current practice.
- Collaborate with open-source developers to create robust libraries for models, tools, and methodologies that support explainability.

RECOMMENDATIONS FOR LAWMAKERS

- Build on existing regulations and require reporting on explainability for high-stakes domains such as finance and medicine.
- Require disclosure of the use of AI systems, ensuring that algorithmic subjects are aware of when and how AI/ML is being used for decisions that affect them.

RECOMMENDATIONS FOR COMPANIES

- Compare the recommendations for counterfactuals to observed data to evaluate their accuracy and effectiveness.
- Test and validate any methods for explainability that they intend to implement, and do not trust methods at face value.
- Build rigorous and automatic evaluation structures into AI pipelines.
- Prioritize the implementation of interpretable methods.

RECOMMENDATIONS FOR RESEARCHERS

- Collaborate with applied AI/ML practitioners to progress counterfactual explanations research.
- Place higher emphasis on safety and explainability in AI research.
- Communicate the risks in using explainability methods developed through your research.

1. Introduction

For decades, AI has promised to revolutionize all parts of public and private life. Despite years of research and planning, the “AI Revolution” has presented novel challenges related to fairness, equity, and transparency. Central among these is a struggle to understand and explain AI systems. As models explode in size and complexity, it becomes more difficult for humans to parse out the processes involved in an AI model’s decision-making.

Understanding AI is particularly difficult for algorithmic subjects, those who are affected by the outcome of an AI-based decision but may have little to no education in AI. Current practice for explaining AI decisions varies from providing broad reasoning to providing no explanation at all. However, as AI reaches further into sensitive and high-stakes domains such as finance and healthcare, transparent, subject-focused methods are needed.

Counterfactual explanations (CFEs), which recommend changes that can be made to achieve a different outcome, have significant promise in improving transparency and understandability. They can provide information in a format that does not require significant knowledge of AI. In particular, CFEs can give specific recommendations that are directly tied to the decision-making system, while humans often have an imperfect knowledge of how they make decisions.

Still, while CFEs offer some improvements over other explanation methods, they still have significant limitations. Companies, lawmakers, researchers, and regulators must keep these limitations in mind when considering how and when to use CFEs.

2. Counterfactuals and Other Explanation Methods

Historically, predictive models were small and simple enough to be understandable at face value. For example, linear regression shows a straightforward relationship between the predictor **x** and the outcome **y** by taking the form **y = ax + b**. Even after adding more variables, the interpretation remains simple: for every **1/a** unit increase in **x**, there is a 1 unit increase in **y**.

However, in the past few decades, state-of-the-art methods have begun modeling much more complex relationships. Researchers developed new approaches for interpretation and explanation alongside these increasingly opaque models. In his seminal paper describing "random forests," a model used in artificial intelligence and machine learning, Breiman derives a method for interpreting his model, and specifically notes the importance of understanding variable interactions in domains such as medicine.[1] For example, a random forests model could be trained to predict if a patient is likely to suffer from a disease. A doctor looking at the model's prediction will care about *how* the prediction was calculated to ensure that the prediction is reasonable.

Two decades later, the random forests method is one of the most widely used approaches for AI/ML prediction tasks, but disagreement remains on how best to interpret a random forests model. Due to the complexity of such models, methods for explaining them to humans fail to capture important intricacies. This illustrates a common problem in machine learning: modeling methods become popular because they attain high accuracy for specific metrics, while concerns around safety, generalizability, and interpretability are often considered secondarily.

Though interpretability and explainability have not been prioritized, they have been and continue to be active areas of research. However, most explainable AI (XAI) research focuses on *developer-facing* methods, rather than *subject-facing* methods; that is, most XAI methods are targeted at individuals who are training AI models,[2] rather than those whose lives are affected by the outputs of the AI's calculations. Consequently, most XAI methods are not immediately useful to algorithmic subjects, those people whose education, finances, and health may be impacted by AI models and decisions.

[1] Breiman, L. "Random Forests." Machine Learning 45 (2001): 5–32.

[2] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2019). Explainable Machine Learning in Deployment. CoRR, abs/1909.06342. http://arxiv.org/abs/1909.06342

Figure 1 showcases some of the differences between “classical” methods of prediction and more modern methods. From this figure, we can see that “classical” methods are simpler graphically as well as mathematically. As we move toward more modern methods, graphs become more complex, and models are no longer mathematically coherent to humans.

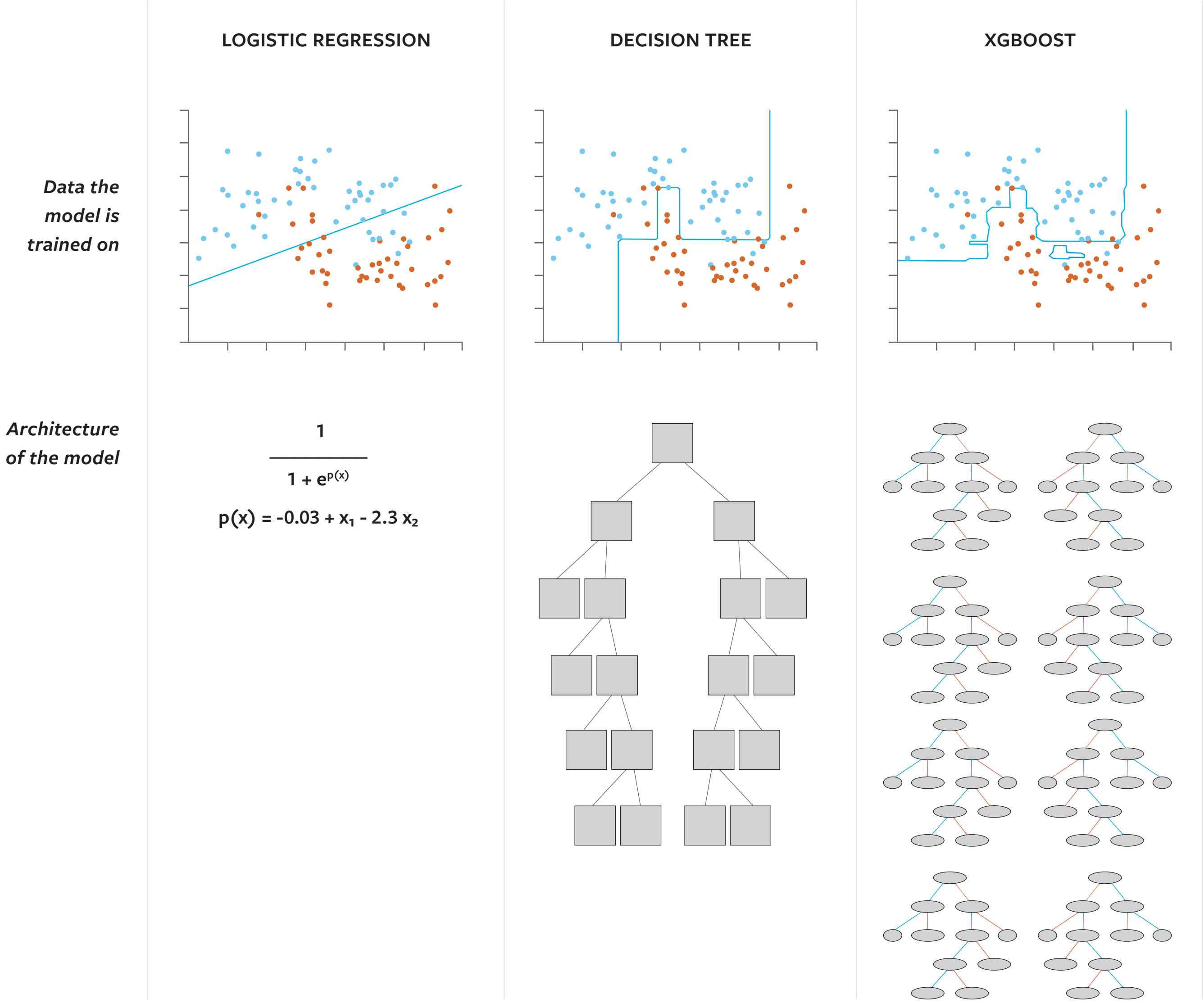
LOGISTIC REGRESSION
DECISION TREE
XGBOOST
Data the model is trained on
Architecture of the model
$\frac{1}{1 + e^{p(x)}}$
$p(x) = -0.03 + x_1 - 2.3\ x_2$

Figure 1: This figure shows three models of increasing complexity (from left to right) trained on the same dataset. The first row displays the data the model is trained on and a line that denotes the prediction structure of the model. One side of the line is a prediction of ‘blue’ while the other is a prediction of ‘red.’ The second row displays the architecture of the model. As is apparent, Logistic Regression is the easiest model to interpret, but the least flexible. The Decision Tree is more complex but still comprehensible, as each split on the tree denotes a decision the model is making. The XGBoost model is incomprehensible, as it is made up of many decision trees; humans could not meaningfully interpret its architecture.

2.1 Explanation Methods

There are several prominent explanation methods, a few of which are described below.

- **Counterfactual explanations** describe how an algorithmic subject can achieve a different decision.[3] For example, if an applicant to a university were to be denied admission based on a decision made by an algorithm, a counterfactual explanation could recommend that the applicant increase their test scores or take additional courses in order to be approved by the algorithm and accepted by the university.
- **SHAP** is based on Shapley values, a concept used in game theory to analyze how coalitions form.[4] In SHAP, predictors of an algorithmic decision are treated as players in a game, and the prediction is treated as the outcome of the game. Over many simulations, predictors are added and removed to determine how much they improve the prediction. More important predictors are assigned higher SHAP values. SHAP can provide helpful clues about what variables a model prioritizes, but does not typically provide individually relevant remedies in the way that counterfactual explanations can.
- **Permutation importance** measures feature importance by shuffling the values within a predictor. We can compare an original model with a model trained on data with a shuffled predictor. The better the original model performs over the shuffled model, the more important the predictor is. Permutation importance can provide helpful insight into the relative importance of different model features, but does not typically provide individually relevant remedies in the way that counterfactual explanations can.
- **Partial dependence plots** show how a model's prediction changes with marginal changes in a predictor. Partial dependence plots disentangle a predictor from the interactions that predictor may be involved in. For example, a model may be trained to predict height based on age and weight. A partial dependence plot for age would show an increase during the first 15-20 years followed by a leveling off. Partial dependence plots can explain how a model makes predictions on one axis, but it is difficult to combine partial dependence plots across variables. Therefore, partial dependence plots are not typically useful for explaining individual decisions.

[3] Wachter, Sandra, Brent Mittelstadt, and Chris Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR" (October 6, 2017). *Harvard Journal of Law & Technology*, 31 (2), 2018. Available at SSRN: https://ssrn.com/abstract=3063289 or http://dx.doi.org/10.2139/ssrn.3063289

[4] Lundberg, Scott M and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems 30*, (2017) : 4765-4774

3. Promises and Limitations of Counterfactual Explanations

3.1. WHAT MAKES A GOOD EXPLANATION?

Although it is widely agreed that explainability is a desirable characteristic of AI systems, there has been rigorous debate on what exactly a “good” explanation is.[5] There are a variety of mathematical, social, and practical considerations to keep in mind.

- **Sparsity**: The “sparsity” of an explanation refers to the number of features that are changed. A sparse explanation would recommend changing a small number of features relative to the total number of mutable features. Consequently, a sparser explanation is a shorter explanation. Sparsity helps ensure that an explanation is useful for subjects in understanding their decision and seeking recourse. Sparsity also enables companies to support customers more readily; instead of walking through all possible options, support services can start with options that are most relevant to the customer at hand. For example, suppose a customer is denied a line of credit. A recommendation to increase their credit score by 50 points would be sparse. In comparison, a recommendation to increase their credit score by 25 points, decrease outstanding debt by $1000, and increase savings by $1500 would be dense. (See Figure 2.)
- **Relevance**: Relevance describes the strength of the connection between the explanation, the algorithmic subject, and the desired outcome. The explanation should only describe variables that are relevant to the decision. For example, an explanation for a diabetes diagnosis should not reference outstanding debt. Relevance ensures that the explanation relates to the underlying circumstance affecting the decision, rather than variables that are not pertinent to the decision.
- **Fairness**: Fairness requires that explanations are of similar quality for all individuals. The exact considerations for fairness depend on the domain, but may include ensuring that explanations are equally accurate across groups and do not require extra effort from some groups. As with explainability, there is active debate about the best definition of “fairness” in AI.[6]

[5] Newman, J. ”Explainability won’t save AI”. TechStream (2021).

[6] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. National Institute of Standards and Technology. https://doi.org/10.6028/nist.sp.1270

- **Actionability:** Actionability requires that an explanation suggests changes that are achievable for the algorithmic subject. In other words, the subject should know from the explanation what they could do to impact the model's decision in the future. Terms such as "feasibility" and "plausibility" are often used to describe this characteristic.

All of these factors must be balanced to achieve what we consider a good explanation. Since AI/ML models are mathematical objects, the mathematical definitions of the above concepts are important to investigate, but we must also stay anchored to the social reasoning behind these definitions.

Explanation Type	Income	Debt	Length of employment	Current unpaid loans	Credit Score
Dense	+20,000	-1,000	+3	-1	+50
Sparse	+30,000	-3,000	—	—	—

Figure 2. This table shows two different explanations in the context of consumer loans; one dense and one sparse. The entries in the table show how much change is required for an applicant to receive a line of credit.

3.1 BENEFITS

With the above criteria in mind, it becomes apparent why counterfactual explanations are considered a substantial improvement in explanations for algorithmic subjects.

- Counterfactuals are **intelligible** to a wide swath of people and complement natural human thought processes. Counterfactuals answer "What if?" questions that a subject may have and suggest specific changes that can be made to achieve a better outcome.
- **Auditing** counterfactual explanations is straightforward; one must simply look for any counterfactuals that recommend impossible or illegal changes. If a counterfactual were to recommend changing one's race or gender, for example, it would signal that the explanation method or underlying model is not appropriate for the task.
- **Privacy** is maintained with counterfactuals, in that the underlying model is not exposed to the public. While there still is some concern that counterfactuals can be used to discover a model that is private (or may contain private data), other explanation methods reveal the model almost entirely. Counterfactual explanations only reveal a single data point per subject, rather than the entire model.

- **Agency:** Counterfactuals could give significantly more agency to subjects than current processes. In the example of a loan application, for example, credit companies are currently required to give some reasoning for denying a loan, but they have no obligation to provide specific guidance on what changes an applicant could make for their application to be approved.

3.3 LIMITATIONS

Counterfactuals have strong theoretical potential but significant practical limitations. The most significant issue is that it is difficult to produce counterfactuals that satisfy all of the characteristics desired in an explanation. Counterfactuals also run into the problem of causal inference; how do you determine how an intervention will affect someone prior to intervening? This question flows through all issues discussed below.

- **Extrapolation**, or making decisions outside the domain of prior data, is central to several of the core issues with counterfactuals. Counterfactuals are by their nature extrapolating; they attempt to make predictions about the future of an individual. To get around this limitation, many counterfactual explanation methods assume that individuals are interchangeable, allowing one individual's experience to be applicable to another's. In reality, it is rarely the case that two individuals are exactly the same. Improper extrapolation can lead to non-actionable and unfair recommendations by ignoring the complex relationships between variables, entrenching current systemic problems, or simply recommending actions that are impossible. For example, a counterfactual may recommend that an individual increase their length of employment by two years. While possible, this recommendation requires an individual to wait for years to seek recourse, which may not be practical.
- **Computation** for counterfactuals can be extremely expensive, especially as model and data sizes increase. To obtain counterfactuals, we must search through the potential ways a subject can change. The amount we must search increases exponentially with each additional variable in our model. As a result, searching for counterfactuals is computationally expensive, potentially taking more time than training the underlying model. Though there are methods to speed up counterfactuals, they produce worse counterfactuals.
- **Manipulation** is a concern for both companies and algorithmic subjects. Companies may fear that releasing explanations will allow their decision-making processes to be gamed more easily. Subjects may be concerned that companies can manipulate counterfactuals to pick winners and losers.

- **Hallucinations**: Counterfactual explanation methods commonly “hallucinate” recommendations, that is, they recommend actions that are not feasible.

3.4 FUTURE DIRECTIONS

Moving forward, it will be crucial that methods for generating counterfactuals are better at extrapolation. While computation and manipulation are important axes for explainability methods, poor extrapolation makes the fundamental goal of explanation impossible, even with infinite computational capacity and the best intentions. Improving extrapolation, which helps counterfactual explanation methods to address individual realities and the complex relationships between variables, is one of the recommendations for XAI researchers described below. Other recommendations are intended to help provide guidance to different stakeholders about how to improve practical understanding of AI systems, despite the imperfect methodologies currently available.

4. Recommendations

Based on the analysis of benefits and limitations, we provide recommendations for government regulators and lawmakers, private companies, and academic researchers who may be working on counterfactual explanations. In some cases, the following recommendations also extend to explainability and ML/AI research more broadly.

4.1 RECOMMENDATIONS FOR REGULATORS

- **Do not require counterfactual explanations as they exist currently, due to deficiencies in current methodologies.**
  Counterfactuals have significant potential but also concerning drawbacks. Given this, regulators should be open to their use while keeping a close eye on their implementation. Requiring counterfactuals, especially at such an early stage in their development, would lead to poor outcomes for both companies and consumers. More broadly, regulators should not require explanations from black box models before verifying that these explanations are sufficient. In cases where explanations are needed or legally required, regulators should caution against the use of black box models and take enforcement action when necessary.
- **Lead the development of frameworks for explainability based on domain knowledge and department needs, rather than relying on current practice.**
  AI/ML practitioners use a variety of methods for explainability in practice, but popular methods may not necessarily align with the principles of explainability within a certain domain. Regulators should support the development of frameworks for explainability based jointly on principles of explainability and on domain knowledge. Government bodies ranging from the Food and Drug Administration (FDA) to the Consumer Financial Protection Bureau (CFPB) have grappled with how to address the increasing use of AI/ML in their respective domains. While there are some similarities across domains, regulatory requirements vary substantially. With this in mind, regulators should build frameworks for explainability based on the fundamental purpose of the regulation. That is, requirements for explainability should be based on the regulation itself, rather than being based on what is easy to satisfy with existing AI/ML models.

For example, the CFPB has recently shown some concern with AI/ML models used in finance. However, they have issued little guidance on what a “good” explanation should contain. This makes it difficult for companies to know if they are at risk of being fined or charged. To help rectify this, the CFPB could release a document detailing what does (and does not) violate regulations. Moreover, the CFPB should treat AI/ML as a tool within consumer finance rather than an exception to the regulatory regime. As such, the normal rules still apply, and regulations pertaining to disclosure are still relevant to AI.

4.2 RECOMMENDATIONS FOR LAWMAKERS

- **Build on existing regulations and require reporting in high-stakes domains such as finance and medicine.**
  Sensitive and high-stakes domains already have existing regulatory frameworks that protect citizens from discrimination, manipulation, and abuse. Lawmakers should take the spirit of these frameworks and extend them as needed to mitigate the negative effects of AI/ML. In domains that already have reporting requirements, lawmakers should ensure that AI/ML systems are included in regulatory reports. These reports should include information that is already required by existing regulation, as well as information on how models are explained and how models are evaluated internally.
- **Require disclosure of the use of AI systems.**
  As with data usage disclosures, algorithmic subjects should be aware of when and how AI/ML is being used. Since data is most often used for training AI/ML, this is a natural extension of existing regulation. Disclosures should give relevant information to algorithmic subjects in an accessible manner and empower consumers to make decisions about which companies they do business or share data with.

4.3 RECOMMENDATIONS FOR COMPANIES

- **Compare recommendations for counterfactuals to examples from previously seen data.**
  Counterfactual explanation methods commonly "hallucinate" recommendations, that is, they recommend actions that are not feasible. One way to combat this is to compare counterfactuals to observed data. This aids in deciding if the recommendation made by a counterfactual is feasible and prevents decision-makers, such as private companies, from giving nonsensical recommendations.

- **Test and validate any methods you intend to implement; do not trust methods at face value.**
  Companies are responsible for the effects of their products on consumers, and they must take every precaution to ensure that AI/ML models do no harm. The complexity of a model does not excuse the responsibility to avoid harm. With this in mind, companies should test any methods they intend to implement and ensure that they comply with relevant regulations. Importantly, companies should not assume that model performance in another area immediately translates to performance in their own business. As such, every new method should be validated before being used in practice.
- **Build rigorous and automatic evaluation structures into AI pipelines.**
  After deciding to implement a method or model, it is crucial to build rigorous evaluation pipelines. These pipelines should be based on domain-relevant and regulation-compliant metrics. Additionally, evaluation pipelines should provide insight into how a model arrives at decisions and give an opportunity to stop the model if something goes wrong. Pipelines should be set up to perform evaluations automatically to speed up detection of potential harms.
- **Prioritize implementation of interpretable methods.**
  Current practice in machine learning is to train a complex, uninterpretable model without investigating a simpler, more interpretable model. It is assumed that complex models perform better than interpretable models, but this is often not the case. Excluding some particularly difficult problems, such as vision and language modeling, interpretable models often perform just as well as black box models. Moreover, interpretable models are much easier to test, validate, edit, and explain than black box models, easing the process of model improvement. While interpretable models do not get rid of the complexities in evaluation, they do make it easier to combat harm by showing how a model arrives at its decisions. Interpretable models are therefore easier to catch and debug before they are deployed.

4.4 RECOMMENDATIONS FOR RESEARCHERS

- **Collaborate with applied AI/ML practitioners to advance research on counterfactual explanations.**
  Explainability researchers have put substantial emphasis on the feasibility of counterfactual explanations in the past few years, and significant progress has been made toward methodologies that are safer for broad use. However, counterfactuals are not often used in practice, partially due to the gap between methodological research and

applied practice. To bridge this gap, researchers should develop collaborations with applied practitioners. These collaborations will assist in satisfying the needs of people who would use counterfactual explanations in practice.

- **Place higher emphasis on safety and explainability in AI research.**
  The past decade has seen significant growth in the emphasis on safety and explainability in AI/ML research, yet the work is still in the early stages. Every paper that describes a new model—and every open-source code release—should consider and discuss risk. This is not to say that all models must be explainable, but researchers should consider paths to explaining models in conjunction with developing them, similar to Breiman and Random Forests.
- **Communicate the risks in using explainability methods developed through your research.**
  The culture of research disincentivizes communicating limitations, which contributes to dangerous applications of AI/ML. When releasing a model or method, researchers should be transparent about what the limitations are and how they may be addressed, if at all. However, this is a systemic change that has to be made to publication and grant approval structures, rather than something individual researchers necessarily have the power to do. Fortunately, there is a growing community of researchers focused on prioritizing safety and transparency when releasing new models and methods.[7]
  One promising solution for increasing transparency is model cards. Model cards can be used to explain the entire lifecycle of an AI model in a way that is approachable to everyone. Additionally, model cards can elucidate limitations and potential problems before the model is implemented in practice.[8]

[7] Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations.

[8] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019, January). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3287560.3287596

Conclusion

The rise of AI/ML has led to a growing need for explainability and transparency from what are often opaque systems. Counterfactual explanations are a promising tool in the pursuit of explainable AI, but CFEs have significant limitations. Regulators, legislators, private companies, and researchers all have a role to play in improving counterfactuals and increasing explainability in AI/ML more generally.

Acknowledgments

First, I’d like to acknowledge the UC Berkeley AI Policy Hub for graciously funding this work through the AI Security Initiative at the Center for Long-Term Cybersecurity (CLTC) and the CITRIS Policy Lab at the Center for Information Technology Research in the Interest of Society and the Banatao Institute (CITRIS). Moreover, I’d like to thank Brandie Nonnecke, Jessica Newman, and the other AI Policy Hub Fellows for their guidance, perspective, and mentorship throughout my time in the AI Policy Hub, along with invaluable feedback on this draft. My PhD advisor, Giles Hooker, was central in developing my technical expertise.

About the Author

ALEXANDER ASEMOTA is a PhD candidate in statistics at UC Berkeley. His research focuses on explainability, fairness, and auditing in machine learning. He was a fellow in the AI Policy Hub during 2022–2023, and he currently is a trainee in the Computational Research for Equity in the Legal System (CRELS) program. Alexander is a graduate of Howard University.

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

Center for Long-Term Cybersecurity | cltc.berkeley.edu | @CLTCBerkeley