The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons

Solon Barocas [1] **Andrew D. Selbst** [2] **Manish Raghavan** [3]

Abstract

Counterfactual explanations are gaining prominence within technical, legal, and business circles as a way to explain the decisions of a machine learning model. These explanations share a trait with the long-established "principal reason" explanations required by U.S. credit laws: they both explain a decision by highlighting a set of features deemed most relevant—and withholding others.

These "feature-highlighting explanations" have several desirable properties: They place no constraints on model complexity, do not require model disclosure, detail what needed to be different to achieve a different decision, and seem to automate compliance with the law. But they are far more complex and subjective than they appear.

In this paper, we demonstrate that the utility of feature-highlighting explanations relies on a number of easily overlooked assumptions: that the recommended change in feature values clearly maps to real-world actions, that features can be made commensurate by looking only at the distribution of the training data, that features are only relevant to the decision at hand, and that the underlying model is monotonic.

While new research suggests several ways that feature-highlighting explanations can work around some of the problems that we identify, the disconnect between features in the model and actions in the real world—and the subjective choices necessary to compensate for this—must be understood before these techniques can be usefully implemented.

[1]Microsoft Research and Cornell University [2]University of California, Los Angeles [3]Cornell University. Correspondence to: Solon Barocas <sbarocas@cornell.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

1. Introduction

Explanations are increasingly seen as a way to enhance the autonomy of people subject to algorithmic decision making. Advocates assert that they allow people to navigate the rules that govern their lives, help people recognize when they should contest decisions or object to the decision-making process, and facilitate direct oversight and regulation of algorithms (Wachter et al., 2017b; Selbst & Barocas, 2018).

In this paper, we examine two related approaches to explanation: the counterfactual explanations that have been explored in recent computer science research and which are gaining traction in practice and the "principal reason" approach drawn from U.S. credit laws. These two approaches are what we call "feature-highlighting explanations"—approaches that draw decision subjects' attention to a limited set of features that are said to "explain" the decision.

There are at least four reasons for the growing popularity of feature-highlighting explanations. First, this approach appears to allow practitioners to abandon any constraints on model complexity—a constraint often seen as a barrier to improved model performance. Second, it allows institutions to avoid disclosing models in their entirety, thereby protecting trade secrets and businesses' other proprietary interests, while limiting decision subjects' ability to game the model. Third, the approach promises a concrete justification for a decision or precise instructions for achieving a different outcome. Fourth, it allows firms to automate the difficult task of generating explanations for a model's decisions—explanations that are also thought to comply with legal requirements both in the United States and Europe.

Generating feature-highlighting explanations is far from straightforward, however, and requires decision makers to make many consequential and subjective choices along the way. In this paper, we demonstrate that the promised utility of feature-highlighting explanations rests on four key assumptions, easily overlooked and rarely justified: (1) that a change in feature value clearly maps to an action in the real world; (2) that features can be made commensurate by looking only at the distribution of feature values in the

training data; (3) that explanations can be offered without regard to decision making in other areas of people's lives; and (4) that the underlying model is monotonic.

2. What are feature-highlighting explanations?

We define a feature-highlighting explanation as an explanation that seeks to educate the decision subject by pointing to specific features in the model that matter to the individual decision, where each type of feature-highlighting explanation may define "matter" differently. The two types we discuss here are counterfactual and principal reason explanations, though there are others (Ribeiro et al., 2016; Lou et al., 2012; Dhurandhar et al., 2018). Counterfactual explanations, in particular, have begun to attract the interest of businesses, regulators, and legal scholars, with many converging on the belief that such explanations are the preferred approach to explaining machine learning models and their decisions. Principal reason explanations are well established in U.S. credit laws, with various businesses having well developed procedures for generating and issuing so-called "adverse action notices" (AANs)—legally required explanations for adverse decisions. Both methods aim to produce explanations of a *particular* decision by highlighting factors deemed most deserving of the decision subject's attention; they do not aim to explain the decision-making process in general. This section will describe both approaches and how they relate to each other.

2.1. Counterfactual explanations

Recent proposals from computer scientists have focused on generating counterfactual explanations for the decisions of a machine learning model (Martens & Provost, 2014; Wachter et al., 2017b; Ustun et al., 2019; Mothilal et al., 2020; Karimi et al., 2019; Hendricks et al., 2018; Grath et al., 2018). The goal of counterfactual explanations is to provide actionable guidance—to explain how things could have been different and provide a concrete set of steps a decision subject might take to achieve a different outcome in the future. Counterfactual explanations are generated by identifying the features that, if minimally changed, would alter the output of the model.

In particular, an emerging theme in the computer science literature is to frame the search for such features as an optimization problem, seeking to find the "nearest" hypothetical point that is classified differently from the point currently in question (Wachter et al., 2017b; Mothilal et al., 2020; Russell, 2019; Karimi et al., 2019; Ustun et al., 2019). In casting the search for counterfactual explanations as an optimization problem, a key challenge is to define a notion of distance. Different features are rarely directly comparable because they are represented on numerical scales that do not meaningfully map onto one another. We discuss this challenge more in Section 3.2.

Wachter et al. have also argued that counterfactual explanations could satisfy the explanation requirements of the E.U.'s General Data Protection Regulation (GDPR) (Wachter et al., 2017b). Over the last several years, lawyers and legal scholars have debated whether certain provisions of the GDPR create a right to an explanation of algorithmic decisions, and, if it exists, whether and when it requires an explanation of specific decisions or the model (Kaminski, 2019; Selbst & Powles, 2017; Brkan, 2019; Wachter et al., 2017a; Edwards & Veale, 2017; Casey et al., 2019; Mendoza & Bygrave, 2017; Malgieri & Comandé, 2017). The official interpretation of the Article 29 Working Party—a government body charged with creating official interpretations of European data protection law—has concluded that the GDPR requires, at a minimum, explanations of specific decisions (Article 29 Data Protection Working Party). Thus, part of the rationale to employ counterfactual explanations is to satisfy the legal requirements of the GDPR.

2.2. Principal reason explanations

The other type of feature-highlighting explanation is what we call a principal reason explanation. The principal reason approach has a long history in the United States, where the Fair Credit Reporting Act (FCRA) (Fair Credit Reporting Act, Public Law 91-508), Equal Credit Opportunity Act (ECOA) (Equal Credit Opportunity Act, Public Law 93-495), and Regulation B (Regulation B) require creditors—and others using credit information—to provide consumers with reasons explaining their adverse decisions (e.g., consumers being given a subprime interest rate, refused credit outright, or denied a job based on information from their credit file, etc.) (Selbst & Barocas, 2018). Under ECOA and Regulation B, these decision makers are required to issue AANs to such consumers; under FCRA, consumers are given a list of "key factors." These notices must include a statement of no more than four "specific reasons" for the adverse decision (Fair Credit Reporting Act, Public Law 91-508; Regulation B).[1] A Sample Form in the Appendix to the regulation offers a non-exhaustive list of acceptable reasons, such as "income insufficient for amount of credit requested," "unable to verify income," "length of employment," "poor credit performance with us," "bankruptcy," and "no credit file" (Regulation B, Appx. C, (Sample Form)). Under the regulation, "no factor that was a *principal reason* for adverse action may be excluded from disclosure, [and t]he creditor must disclose the *actual reasons* for denial" (Regulation B,

[1]The number four is not a hard limit under Regulation B, as it is under FCRA, but it is observed in practice.

§ 1002.9(b)(2), emphasis added).

What counts as a principal reason is not well-defined in either the statutes or regulation. The legislative history of ECOA indicates that consumer education is a primary goal. This would seem to suggest that counterfactual explanations, as currently conceived, would serve the intended purpose of AANs. And indeed, some scholars have suggested as much (Ustun et al., 2019). But this very ambiguity also demonstrates that principal reasons are satisfied by a broader array of possible feature-highlighting explanations. For example, the Official Staff Interpretation to Regulation B, originally published in 1985 (Federal Register), suggests two ways creditors can generate principal reasons:

> One method is to identify the factors for which the applicant's score fell furthest below the average score for each of those factors achieved by applicants whose total score was at or slightly above the minimum passing score. Another method is to identify the factors for which the applicant's score fell furthest below the average score for each of those factors achieved by all applicants (Regulation B, Supplement I).

Note that neither approach uses the decision boundary as the relevant point of comparison. Instead, they compare the value of applicants' features to the average value of these features for the credit-receiving or general population, in an attempt to surface the dimensions along which the applicant is most deficient.[2]

2.3. Highlighting subsets of features in the service of autonomy

For our purposes, counterfactual and principal reason explanations have one crucial thing in common: neither involves disclosing the model in its entirety. They focus, instead, on highlighting a limited set of features that are deemed most deserving of a decision subject's attention. By design, they do not provide an exhaustive inventory of all the features that a model considers. In practice, learned models can consider a very large set of features, and an explanation that suggests changes to each of those features would be overwhelming. As a result, both the law (in the form of principal reasons) and the emerging technical literature (in the form of counterfactual explanations) seek to produce "sparse" explanations that present the decision subject with only a small subset of features (Wachter et al., 2017b).

When opting for this style of explanation, there is no natural way to choose between the principal reason and counterfactual approaches. Yet rarely is the choice to use one method over another discussed explicitly or even recognized as a choice in the first place. These methods produce different explanations and serve fundamentally different goals.

Focusing on features that are furthest from the average value of the features in the credit-receiving or general population casts the problem of identifying principal reasons as one of identifying extreme deficiencies that would seem to rule out the applicant completely, rather than near-misses that applicants might readily address before applying for credit again in the future. While the former may strike us as a less attractive or sensible approach to explanation, there may be good reason to favor an explanation that makes clear the features that were held against an applicant. With the latter approach, while the applicant might receive helpful advice, she might not learn that other features were viewed by the model as crucial marks against her.

Principal reason explanations treat importance in terms of procedural justice: to respect the autonomy of a decision subject, the decision subject deserves to know which factors dominated the decision (Tyler, 2006). In counterfactual explanations, respect for autonomy means that decision subjects need to know how choices affect outcomes, and thus how they can take actions that will most effectively serve their interests in the future. The former operates more like a justification for a decision—a rationale, with little immediate concern for recourse—whereas the latter serves a more practical purpose—providing explicit guidance for achieving a different decision in the future. Crucially, both styles of explanations can be educational, even if they differ in how easily decision subjects can act on them.

In keeping with these differences, the principal reasons offered by creditors tend to be vaguer ("Income insufficient for amount of credit requested"), while counterfactual explanations aim for precision ("Had you earned $5,000 more, your request for credit would have been approved"). In many cases, principal reason explanations do not even disclose the magnitude, let alone the direction, of change that would be necessary to achieve a different outcome, while such details are inherent to counterfactual explanations. In fact, the official interpretation of the regulation that requires creditors to issue AANs notes that creditors "need not describe how or why a factor adversely affected an applicant. For example, the notice may say 'length of residence' rather than 'too short a period of residence'" (Regulation B, § 1002.9(b)(2) (Official Interpretation)).

[2] Though they are written into the regulation, it is not clear that firms actually use these methods to generate principal reasons.

3. Feature-highlighting explanations in practice

There are several hidden assumptions behind the belief that feature-highlighting explanations will be useful for decision subjects. In this section, we identify four such assumptions, explain why they might not be valid, and explore the consequences of that realization.

3.1. Features do not clearly map to actions

Feature-highlighting explanations often assume a clear and direct translation from suggested changes in feature values to actions in the real world. In many cases, this is a reasonable assumption: instructing someone to reduce their total lines of credit maps onto the obvious action of cancelling a credit card or fully repaying—and thus dispensing with—a loan. In most of the contexts in which existing scholarship considers the challenge of explaining the decisions of a machine learning model, there is a clear correspondence between the feature values that one is told to change and the actions that one would take to achieve those changes. And yet, in many cases, the actions that a decision subject can take to affect those features may not line up with the features themselves. For example, a recommendation that someone increase his income can lead a person to take one of several actions: he can seek a new job, ask for a raise, or take on more hours. As Figure 1 illustrates, these actions are not as simple as "increase income" or "increase length of employment."

Highlighting certain features as those that need to change to obtain a different decision implicitly relies on the belief that everything else can be held constant while making these changes. In reality, actions may affect multiple features simultaneously because actions do not line up with discrete features. Changes in the value of one feature may also affect the value of another feature, if the two features interact. As Figure 1 demonstrates, whether one increases his income by finding a higher-paying job or waiting for his performance review to get his raise, the action will affect both income *and* length of employment, a separate feature in the model. In the case of a job change, length of employment will be negatively affected. Thus, increasing income may not be enough to get credit, which is why point (1) is on the left of the decision boundary. In the case of waiting for a raise, a smaller increase in income might be needed than the explanation would dictate, because length of employment increases at the same time. This is why point (2) is on the right side of the decision boundary, despite not increasing income as much as the explanation suggests. When considered in the context of feasible real-world actions, it becomes clear that features may not only fail to line up with actions, but may not be *independent*: the action necessary to change one feature could impact others.

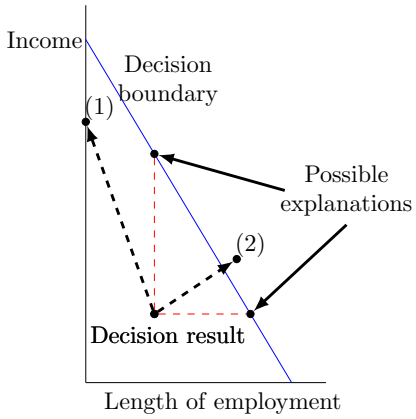

Figure 1. A decision based on two features—income and length of employment—will be explained by reference to one of the features, either the shortest or longest distance from the boundary. But the explanations do not map to the decision subject's possible actions that can affect them. Point (1) represents getting a higher-paying job, and point (2) represents waiting for a raise.

So far, this example has assumed a relatively clear relationship between actions and features. In some cases, though, we may lack the necessary domain knowledge or fail to possess an understanding of the relevant causal mechanism to relate specific actions in the real world to predictable changes in feature values. Under such circumstances, we may struggle to identify the actions that would cause a feature value to change—or change in the way we want.

Recent work on "actionable" explanations has focused on a particular version of this problem: when there are *no* actions that a decision subject can take to change the value of certain features. Scholars have argued that we should avoid explanations that tell people to make changes that are impossible, placing the burden on decision makers to give advice that is sensitive to the actual steps that decision subjects would need to take to achieve the change in feature values (Ustun et al., 2019; Mothilal et al., 2020). Yet avoiding these potential explanations is really a matter of identifying the lack of any possible causal mechanism in the real world that would have the necessary effect on the value of some feature.

"Gaming" is yet another case of this disconnect between actions and changes in feature values (Bambauer & Zarsky, 2018). When a decision maker instructs someone to change certain features, the decision maker will often assume that the person will take a specific desirable sequence of actions because that is the causal mechanism that the decision maker has in mind for changing the value of these features. But there are often many other ways to change feature values that don't require taking these steps (Klein-

berg & Raghavan, 2019).

Avoiding these problems requires identifying all the relevant dependencies and constraints in advance. This is quite a tall order, not least because the possible actions—or lack thereof—that are available to decisions subjects may not always be self-evident. Insisting that explanations exhibit sensitivity to these dependencies and constraints is analogous to insisting that explanations take into account the causal mechanisms that allow decision subjects to alter the value of specific features (Miller et al., 2019).

3.2. Features cannot be made commensurate by looking only at the distribution of the training data

All feature-highlighting explanations rely on some notion of a distance between the observed values for various features and some reference point, whether the the decision boundary or the average value in the population. Relying on distance requires normalizing features, because there needs to be a shared scale between features in order to meaningfully compare them. For example, as discussed in Section 2.1, an increase in length of employment is not naturally commensurate with an increase in salary. Normalization attempts to capture the fact that salaries may vary on the order of thousands or tens of thousands of dollars, but length of employment (in years, say) varies at a numerically much smaller scale.

Several statistical techniques exist to address this problem, scaling features so as to make them seemingly comparable, and different explanation techniques use different approaches. Following Wachter et al., the literature on counterfactual explanations has mostly converged on a heuristic that finds the Median Absolute Deviation (MAD) under an L1 distance norm (Russell, 2019; Mothilal et al., 2020; Grath et al., 2018). Meanwhile, it is entirely unclear what methods principal reason explanations use—the regulations do not specify and it is never discussed in practice—but the nature of a distance metric requires that *something* be used. Normalization techniques are typically based entirely on the distribution of the data, not some external point of reference.

When examined from the perspective of a decision subject who must take some action in response to these explanations, normalization based simply on the distribution of data is somewhat arbitrary. One decision maker might scale the axes such that increasing income by $5,000 annually is equivalent to an additional year on the job. A competing lender, using different training data, could conclude that $10,000 of income corresponds to one year of work. These lenders might therefore produce different explanations depending on the scaling of attributes. Without an external point of reference to ground these scales, the meaning of the relative difference in feature values is unclear.

What counts as seemingly equivalent features according to the distribution of the training data is not necessarily what the decision subject would view as equivalent. Making features commensurate in this way fails to consider the particular circumstances faced by any given decision subject. Because we are concerned with decision subject's menu of options, the most sensible external referent would be something akin to the cost of making the required change, where cost can imply dollars spent, effort, or time. For counterfactual explanations, those features that involve little *cost* to change, even if they involve considerable change along a normalized numeric scale, may be far more useful to highlight than those that would be costlier to change. If, instead, the preferred approach involves generating principal reasons, features that are costly or impossible to change may be precisely the ones that should be highlighted. Thinking about changes in terms of their real-world cost therefore helps to translate numerical changes in feature values to real-world actions, whether we want to point out either what is easiest or most difficult to change. Some recent work seeks to account for the cost of actually manipulating features in practice by assuming domain knowledge or soliciting user input (Ustun et al., 2019; Grath et al., 2018; Mothilal et al., 2020), but decision subjects may be unable to articulate all of the relevant real-world constraints that would affect the utility of an explanation.

Worse yet, the cost of making certain changes will not be consistent across different people. Changes that might be rather inexpensive for one person to make might be costly for another person to make (Rudin, 2019). Thus, when we use explanations to identify the easiest or most difficult features for someone to change to achieve a different decision from a model, the explanation must be sensitive not only to how these changes involve different costs, but how these costs vary across the population. Different subsets of features may be appropriate for different people with different life circumstances. This complication cuts against the very desirability of these explanations: the idea that we can automatically determine what is easiest or hardest to change.

3.3. Features may be relevant to decision making in multiple domains

Feature-highlighting explanations may interact with facts about a person's life that are invisible to the model. In particular, the supposition of a counterfactual explanation is that it is offering advice about the kinds of changes that it would be rational for a person to make to achieve better results in future decisions. Some commentators and scholars have cautioned that explanations should never encourage people to take actions that are irrational or harmful (Hall et al., 2017; Eubanks, 2018). What they mean more specifically is that there may be some recommendations that are indeed rational if the only goal is to obtain a positive de-

cision from the model, but irrational with respect to other goals in a person's life.

A common-sense example for this proposition is that an explanation should never recommend that a person seek to make less money (e.g. Lipton, 2017). While we believe it unrealistic that an actual credit model would ever (be allowed to) learn such a relationship, the example still holds value. It is self-evident that no one would want to make less money to improve their access to credit. Or consider an example that reverses this dependency: a person contemplating applying to a new job for its superior health insurance is unlikely to remain at his current job because an explanation for a failed credit application told him to increase the value of his length-of-employment feature. In this case, acting on the recommendation would impose an opportunity cost on the consumer by forcing him to forgo benefits in other domains. When other aspects of one's life depend on some of the same features, explanations for how to get the desired outcome in one aspect of your life may conflict with those in another.

We can reason about this the other way around as well. From the point of view of a counterfactual explanation, an applicant might be best off trying to change a number of other features *besides* income. Yet, from the perspective of the applicant, increasing her income might have ancillary benefits in other parts of her life that make this change more attractive—and indeed rational—than those suggested by the explanation. Increasing her income would grant her improved access not only to credit, but to improved quality of life, generally.

In the first case, a change in feature might benefit the decision subject in one domain, while hurting her in others. In the second case, a change in a feature might benefit the decision subject in multiple domains, not just one. These spillovers—both negative and positive—complicate the process of determining which features would be most useful to highlight in an explanation. Ideally, feature-highlighting explanations would allow decision subjects to avoid negative spillovers and identify opportunities for positive spillover. But a decision maker will lack information about the many other goals that a person might have in her life and the features that are relevant in those domains.

This also highlights an additional risk: due to other life goals, decision subjects may change *undisclosed* features unless otherwise instructed. For example, if a counterfactual explanation tells someone to increase her income and lower her debt, but fails to mention that she should not reduce her length of employment, she may have no idea that she should avoid any career change while attempting to address these other issues, stumbling accidentally into point (1) in Figure 1. Indeed, she might not even know that length of employment figured into the credit decision in the first place. Thus, by failing to disclose what a decision subject must *not* change, an explanation may lead her to take an ultimately unsuccessful action.

3.4. Models may be non-monotonic

Feature-highlighting explanations can also be misleading if the model has not been subject to monotonicity constraints, which guarantee that as the value of the features move in the recommended direction, the decision subject's chances of success consistently improves. Without monotinicity constraints, a model might learn complex and even counter-intuitive relationships between certain features and an outcome of interest. For example, a model might learn that people who have spent two to four years at their current job are good candidates for credit, while those who have stayed five or more are not. Likewise, carrying more debt might render applicants less attractive, until they start earning more income, at which point additional debt might make them more attractive.

Decision subjects will not necessarily be able to alter the value of these features through some sudden step change. Instead, they may have to make incremental changes in the direction of the specified value. And despite their best efforts, decision subjects might struggle to hit the specified feature value; their efforts could move the value of these features in the right direction, but ultimately fail to get the decision subject all the way there. Similarly, decision subjects might lack precise control over the value of a feature, making it difficult to avoid overshooting the mark when they take some action. Unless the model exhibits monotonicity with respect to the highlighted features, the decision subject might find herself in a *worse* position as she moves toward the specified value or if she exceeds it.

4. Conclusion

Feature-highlighting explanations have been embraced as a way to help decision makers avoid a number of difficult trade-offs, granting institutions the capacity to provide meaningful and useful explanations of machine-learned models without having to compromise on model performance, while also respecting concerns with trade secrecy, gaming, and legal compliance. Advocates have championed this style of explanation as an elegant way to honor and enhance decision subjects' autonomy even as machine learning models grow in complexity and ubiquity. Yet as we have shown, these explanations lack a connection to the real-world actions required to change features, often promising far more than they will be able to deliver. While considerable work is already underway that tries to address some of these limitation, it is still far from clear when feature-highlighting explanations can be useful to decision makers and decision subjects alike.

References

Article 29 Data Protection Working Party. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, 2017.

Bambauer, J. and Zarsky, T. The algorithm game. *Notre Dame L. Rev.*, 94:1, 2018.

Brkan, M. Do algorithms rule the world? algorithmic decision-making and data protection in the framework of the gdpr and beyond. *International journal of law and information technology*, 27(2):91–121, 2019.

Casey, B., Farhangi, A., and Vogl, R. Rethinking explainable machines: The gdpr's right to explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*, 34:143–188, 2019.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pp. 592–603, 2018.

Edwards, L. and Veale, M. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16: 18–74, 2017.

Equal Credit Opportunity Act, Public Law 93-495. Codified at 15 u.s.c. § 1691, et seq., 1974.

Eubanks, V. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

Fair Credit Reporting Act, Public Law 91-508. Codified at 15 u.s.c. § 1681, et seq., 1970.

Federal Register. 50 fed. reg. 10915, 1985.

Grath, R. M., Costabello, L., Van, C. L., Sweeney, P., Kamiab, F., Shen, Z., and Lecue, F. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*, 2018.

Hall, P., Phan, W., and Ambati, S. Ideas on interpreting machine learning. https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning, 2017.

Hendricks, L. A., Hu, R., Darrell, T., and Akata, Z. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

Kaminski, M. E. The right to explanation, explained. *Berkeley Tech. LJ*, 34:189–218, 2019.

Karimi, A.-H., Barthe, G., Belle, B., and Valera, I. Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2019.

Kleinberg, J. and Raghavan, M. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 825–844. ACM, 2019.

Lipton, Z. The mythos of model interpretability. Panel discussion: Algorithms and Explanations: Modes of Explanation in Machine Learning, NYU Algorithms and Explanations Conference, April 27 2017.

Lou, Y., Caruana, R., and Gehrke, J. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 150–158. ACM, 2012.

Malgieri, G. and Comandé, G. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, 7:243–265, 2017.

Martens, D. and Provost, F. J. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–99, 2014.

Mendoza, I. and Bygrave, L. A. The right not to be subject to automated decisions based on profiling. In *EU Internet Law*, pp. 77–98. Springer, 2017.

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. *arXiv*, pp. arXiv–1910, 2019.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2020.

Regulation B. 12 c.f.r. § 1002 et seq.

Ribeiro, M. T., Singh, S., and Guestrin, C. 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. ACM, 2016.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Russell, C. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 20–28, 2019.

Selbst, A. D. and Barocas, S. The intuitive appeal of explainable machines. *Fordham Law Review*, 87:1085, 2018.

Selbst, A. D. and Powles, J. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017.

Tyler, T. R. *Why people obey the law*. Princeton University Press, 2006.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19. ACM, 2019.

Wachter, S., Mittelstadt, B., and Floridi, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017a.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gpdr. *Harv. Journal of Law & Technology*, 31:841–887, 2017b.