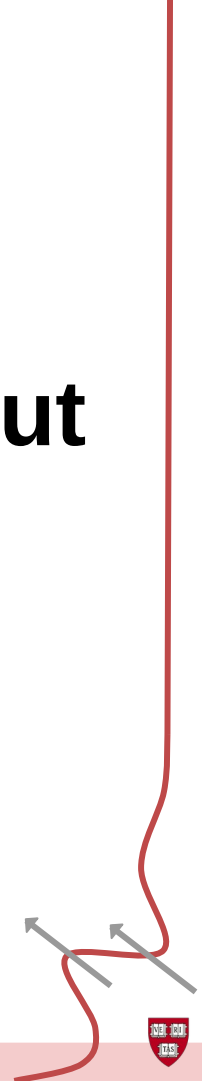
Counterfactual Explanations Without Opening the Black Box

Paper by: Sandra Wachter, Brent Mittelstadt, & Chris Russell
Presented by: Alex, Chelse, & Usha



Introduction + Motivation

- EU General Data Protection Regulation (2018)
  - “the toughest privacy and security law in the world”
  - Article 13-14, regarding automated decision-making: “meaningful information about the logic involved”
  - Recital 71: “the right to obtain an explanation of the decision reached and to challenge the decision”

4 key problems

1. Not legally binding
2. Only applicable in limited cases
3. Explainability is technically very challenging
4. Competing interests of data controllers, subjects, etc

Unconditional Counterfactual Explanations

Authors propose

- 3 aims for explanations
  - Inform and help the subject understand “why”
  - Provide grounds to contest decisions
  - Understand options for recourse
- CFEs
  - Fulfill the above goals
  - Overcome challenges w.r.t. current interpretability work
  - Can bridge the gap between interests of data subjects and data controllers

Do you buy this?

Unconditional Counterfactual Explanations

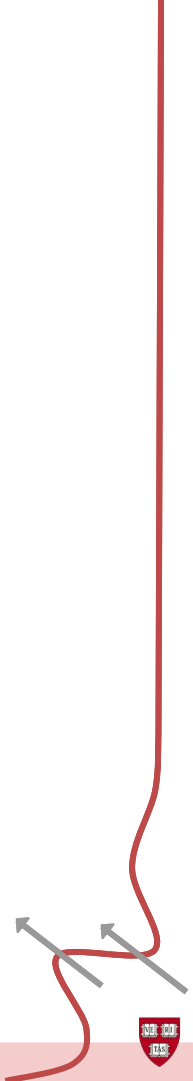
- Definition:
  - A statement of how the world would have to be different for a desirable outcome to occur
  - **Because** of features $\{x_1, x_2\}$, x's outcome was label y_1
  - **If** $x = \{x_1 + \delta_1, x_2 + \delta_2\}$, **then** y_2.
- Example:
  - "You were denied a loan because your annual income was $30000. If your income has been $45000, you would have been offered a loan"

Notes:

- No one **unique** CF
- Need to consider **actionability** (mutability of variables)
- Providing **several diverse** CFEs may be more useful than simply the closest/shortest one

Background + Related Work

- Historic context of knowledge
- Prior explainability work
- Adversarial examples/perturbations
- Causality/Fairness



Background: Historic Context of Knowledge

- In order to know something, it is not enough to simply **believe** that it is true: rather, you must also have a good **reason** for believing it
- If *q* were false, *S* would not believe *p*

Note:

- This statement only describes *S*'s **beliefs**, which might not reflect reality
- This statement can be made without knowledge of the **causal relationship** between *p* and *q*
- Q: who is *S*? The user? The model? Us?

Background: Previous Explanations in AI/ML

- Previously: providing insight into the internal state of an algorithm, human-understandable approximations of the algorithm
- Three-way tradeoff between
  1. quality of approximation
  2. ease of understanding the function
  3. size of the domain for which the approximation is valid
- CFEs:
  - Minimal amount of information
  - Require no understanding of internal logic of model
  - No approximation (although might not always be minimum length)
  - Con: May be overly restrictive

Background: Adversarial Perturbations

- Small changes to a given image can result in the image being assigned to a different class
- Very similar to CFEs but the changes aren't necessarily sparse
- Often not human perceptible
  - Authors propose that this is because the new images lie outside the "real-image" manifold
  - Emphasize the importance of solutions/CFEs being *possible* as well as close
  - Further research into structure of high-D data is structured before CFEs can be useful/reliable in those domains

Background: Causality and Fairness

- Can provide evidence that models/decisions are affected by protected attributes
- If CFEs change one's race, the treatment of that individual is dependent on race

Is the converse true?

Summary of Contributions

- Highlights the difficulties with conveying the inner workings of modern ML algorithms to users
  - Complexity
  - Lack of utility (except for “builders”)
- Introduces an algorithmic approach to counterfactuals
  - Rooted in adversarial machine learning
- Connects counterfactuals to the GDPR
  - Demonstrates advantages over other interpretability methods from a policy perspective

Approach: Counterfactual Optimization

$$\arg\min_{x'} \max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

- Minimize:
  - Squared error with desired (counterfactual) label
  - Distance to perturbed point
- While maximizing weight on squared error term
  - We want to have the prediction change more than we want the point to be close
  - Iteratively: solve for x', then maximize lambda
- Settings:
  - Tabular Data
  - Regression

Approach: Distance Metrics

$$d(x, x') = \sum_{k \in F} \frac{||x_{i,k} - x'_k||_p}{N}$$

Norms

$$|x_{i,k} - x'_k|$$

$$(x_{i,k} - x'_k)^2$$

Normalizing Factors

$$MAD_k := \text{median}_{j \in P}(|X_{j,k} - \text{median}_{l \in P}(X_{l,k})|)$$

$$\text{std}_{j \in P}(x_{j,k})$$

Experimental Results: LSAT

$$d(x_i, x') \propto \sum_{k \in F} (x_{i,k} - x'_k)^2$$

	Unnormalised L2								
	Original Data			Counterfactuals			Counterfactual Hybrid		
score	GPA	LSAT	Race	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.0	39.0	0.3	1.5	38.4	0
0.54	3.7	48.0	0	3.5	47.9	0.9	-1.6	45.9	0
-0.77	3.3	28.0	1	3.5	28.1	-0.3	5.3	28.9	0
-0.83	2.4	28.5	1	2.6	28.6	-0.4	4.8	29.4	0
-0.57	2.7	18.3	0	2.9	18.4	-1.0	8.4	20.6	0

Table 1 - Unnormalized L2

	Normalised L2								
score	GPA	LSAT	Race	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.0	37.0	0.2	3.0	34.0	0
0.54	3.7	48.0	0	3.5	39.5	0.4	3.5	33.1	0
-0.77	3.3	28.0	1	3.5	39.8	0.4	3.4	33.4	0
-0.83	2.4	28.5	1	2.7	37.4	0.2	2.6	35.7	0
-0.57	2.7	18.3	0	2.8	28.1	-0.4	2.9	34.1	0

Table 2 - Normalised L2

Experimental Results: LSAT

$$d(x_i, x') = \sum_{k \in F} \frac{|x_{i,k} - x'_k|}{MAD_k}$$

	Original Data			Normalised L1 Counterfactuals Continuous			Counterfactual Hybrid		
score	GPA	LSAT	Race	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.1	35.0	0.1	3.1	34.0	0
0.54	3.7	48.0	0	3.7	33.5	0.0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	34.4	0.1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	39.3	0.2	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	35.8	0.1	2.7	34.9	0

Table 3 - Normalised L1

Experimental Results: Discussion

- Realistic (or achievable) counterfactual values
  - Clipping and Clamping
- Counterfactual Targets
  - LSAT: Avg. Grade 0
    - Students would want to know how to improve their average grade
  - PIMA: Risk Score of 0.5
    - Patients would want to know how to achieve a lower risk score

Explanations and The GDPR

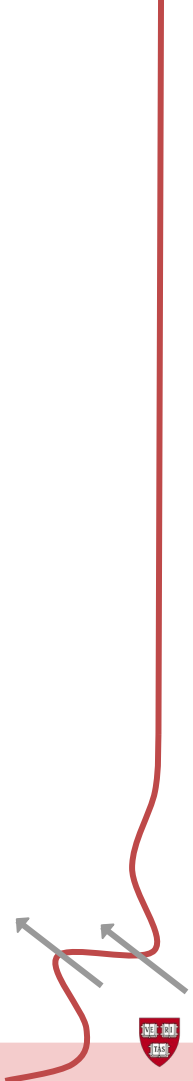
GDPR Requirements: Recital 71

- Implement suitable safeguards against automated decision-making
- Include specific information to the data subject and the right to obtain human intervention
  - To express their point of view
  - To challenge the decision
  - To obtain an *explanation* of the decision reached after such assessment

- Does not require opening the “black box” to explain the internal logic of decision-making systems
- Does not explicitly define requirements for explanations of automated decision-making

Counterfactual Explanations and The GDPR

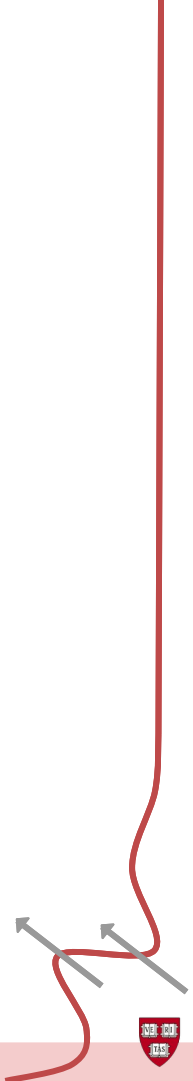
- Legislators wanted to clarify that some type of explanation can voluntarily be offered after a decision has been made
- Many aims for explanations are feasible
- Emphasis of GDPR is on protections and rights for individuals



Advantages of Counterfactual Explanations

- Bypass explaining the *internal workings* of complex machine learning system
- Simple to compute and convey
- Provide information that is both easily digestible and practically useful
  - Understanding the reasons for a decision
  - Contesting decisions
  - Altering future behaviour to receive a preferred outcome

Possible mechanism to meet the explicit requirements and background aims of the GDPR



Broader Possibilities with the Right of Access

Explanations to ***understand*** decisions

Art 15.

- Confirm whether or not personal data used
- Provide information available after a decision has been made
- Avoid disclosing personal data of other data subjects
- Balance interest of the subject and controller
  - Potential to contravene trade secrets or intellectual property rights

Broader Possibilities with the Right of Access

Explanations to ***understand*** decisions

Counterfactuals

- Data of other data subjects or detailed information about the algorithm does not needs to be disclosed
- Disclose only the influence of select external facts and variables on a specific decision
- Less likely to infringe on trade secrets or privacy

Understanding Through Counterfactuals

Explanations to ***understand*** decisions

Art 12(1)

- Requires information to be conveyed in a “concise, transparent, intelligible and easily accessible form”
- CFEs align with this requirement by providing simple “if-then” statements
- CFEs provide greater insight into the data subject’s personal situation as opposed to an overview tailored to a general audience

Minimally burdensome and disruptive technique to understand the rationale of specific decisions

Legal Information Gaps on Contesting Decisions

- Art. 16
  - Data subject has the right to correct inaccurate data used to make a decision, but does not need to be informed of which data the decision depended

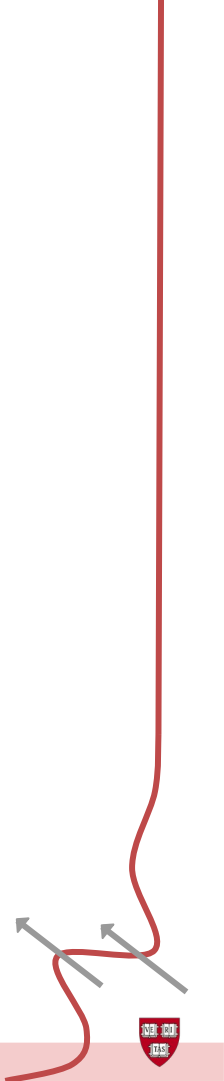
- Art. 22
  - Data subjects do not need to be informed of their right *not* to be subject to an automated decision

Contesting Through Counterfactuals

- Lead to greater protection for the data subject than currently envisioned by the GDPR

- Align with Article 29 Working Party
  - Understanding decisions + knowing legal basis is essential for contesting decisions

- Reduce burden on data subject
  - Understand most influential data (instead of vetting all collected data)
  - Compact way to convey dependencies

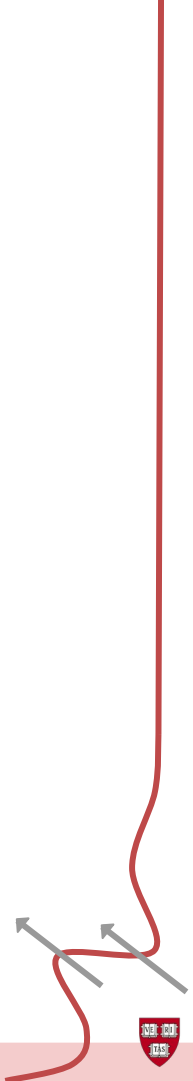


Explanations to Alter Future Decisions

- GDPR
  - Explanations not explicitly mentioned as a guide to altering behavior to receive a desired automated decision

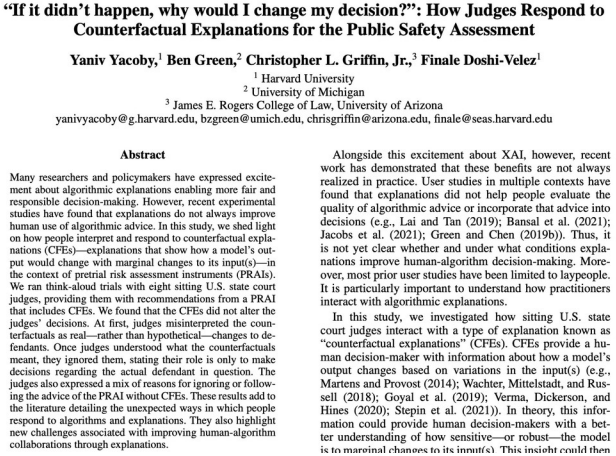
- Article 29 Working Party
  - Provide suggestions on how to improve habits and receive better outcome

- Counterfactual explanations
  - Can address the impact of changes to more than one variable on a model's output at the same time
  - Can be used in a contractual agreement between data controllers and data subjects



Discussion

- Who do these explanations serve?
  - Developers, Data subjects, and Lawmakers
- Do you think they are misinterpreted?
  - Causality
  - Optimizing for sparsity may not reflect reality
- Priors and desired outcomes



"If it didn't happen, why would I change my decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment

Yaniv Yacoby,[1] Ben Green,[2] Christopher L. Griffin, Jr.,[3] Finale Doshi-Velez[1]
[1] Harvard University
[2] University of Michigan
[3] James E. Rogers College of Law, University of Arizona
yanivyacoby@g.harvard.edu, bzgreen@umich.edu, chrisgriffin@arizona.edu, finale@seas.harvard.edu

Abstract

Many researchers and policymakers have expressed excitement about algorithmic explanations enabling more fair and responsible decision-making. However, recent experimental studies have found that explanations do not always improve human use of algorithmic advice. In this study, we shed light on how people interpret and respond to counterfactual explanations (CFEs)—explanations that show how a model's output would change with marginal changes to its input(s)—in the context of pretrial risk assessment instruments (PRAIs). We ran think-aloud trials with eight sitting U.S. state court judges, providing them with recommendations from a PRAI that includes CFEs. We found that the CFEs did not alter the judges' decisions. At first, judges misinterpreted the counterfactuals as real—rather than hypothetical—changes to defendants. Once judges understood what the counterfactuals meant, they ignored them, stating their role is only to make decisions regarding the actual defendant in question. The judges also expressed a mix of reasons for ignoring or following the advice of the PRAI without CFEs. These results add to the literature detailing the unexpected ways in which people respond to algorithms and explanations. They also highlight new challenges associated with improving human-algorithm collaborations through explanations.

Alongside this excitement about XAI, however, recent work has demonstrated that these benefits are not always realized in practice. User studies in multiple contexts have found that explanations did not help people evaluate the quality of algorithmic advice or incorporate that advice into decisions (e.g., Lai and Tan (2019); Bansal et al. (2021); Jacobs et al. (2021); Green and Chen (2019b)). Thus, it is not yet clear whether and under what conditions explanations improve human-algorithm decision-making. Moreover, most prior user studies have been limited to laypeople. It is particularly important to understand how practitioners interact with algorithmic explanations.

In this study, we investigated how sitting U.S. state court judges interact with a type of explanation known as "counterfactual explanations" (CFEs). CFEs provide a human decision-maker with information about how a model's output changes based on variations in the input(s) (e.g., Martens and Provost (2014); Wachter, Mittelstadt, and Russell (2018); Goyal et al. (2019); Verma, Dickerson, and Hines (2020); Stepin et al. (2021)). In theory, this information could provide human decision-makers with a better understanding of how sensitive—or robust—the model is to marginal changes to its input(s). This insight could then