One Explanation to Rule them All — Ensemble Consistent Explanations

André Artelt[1,2] , **Stelios Vrachimis**[2] , **Demetrios Eliades**[2] , **Marios Polycarpou**[2] , **Barbara Hammer**[1]

[1]Bielefeld University , [2]University of Cyprus

{artelt.andre, vrachimis.stelios, eliades.demetrios, mpolycar}@ucy.ac.cy , bhammer@techfak.de

Abstract

Transparency is a major requirement of modern AI based decision making systems deployed in real world. A popular approach for achieving transparency is by means of explanations. A wide variety of different explanations have been proposed for single decision making systems. In practice it is often the case to have a set (i.e. ensemble) of decisions that are used instead of a single decision only, in particular in complex systems. Unfortunately, explanation methods for single decision making systems are not easily applicable to ensembles – i.e. they would yield an ensemble of individual explanations which are not necessarily consistent, hence less useful and more difficult to understand than a single consistent explanation of all observed phenomena. We propose a novel concept for consistently explaining an ensemble of decisions locally with a single explanation – we introduce a formal concept, as well as a specific implementation using counterfactual explanations.

1 Introduction

Nowadays, transparency is a natural and broadly accepted requirement [parliament and council, 2016] of AI based decision making systems. Usually, transparency is realized by providing explanations of the system's behavior [Samek *et al.*, 2017; Guidotti *et al.*, 2018]. While it is not yet completely understood what exactly makes up a "good" explanation [Doshi-Velez and Kim, 2017; Offert, 2017], a lot of different explanation methodologies [Molnar, 2019] such as feature relevances/importances [Fisher *et al.*, 2018] and example based methods [Aamodt and Plaza., 1994] like contrasting explanations [Wachter *et al.*, 2017; Verma *et al.*, 2020] have been developed. Furthermore, explanation methods can be classified into global and local methods [Molnar, 2019]: While global methods try to explain the entire system, local methods explain the system for a particular input or small region of input samples only.

To the best of our knowledge, existing methods focus on single decisions only – i.e. they provide explanations of a single model and (if local) applied to a single input. In scenarios related to complex systems, we face a set or ensemble of decisions instead of a single function that must be explained. Examples are systems which expand in space – i.e. spatial models connected to network data – or in time – i.e. decisions within time-dependent dynamic systems. A highly relevant domain in the real world are water distribution networks where various quantities of interest in different regions of the network are monitored[1] by different systems. These are typically based on the shared sensor data [Perez *et al.*, 2014; Sanz *et al.*, 2016], hence the monitoring system is given by an ensemble of decisions based on shared features – we will address this application domain in Section 4.

In this work we propose a formalization how to provide a consistent explanation for an ensemble of decisions or models – i.e. computing one single explanation instead of a separate and possibly disconnected explanation for each model. More specifically, our contributions are as follows:

1. We introduce the abstract concept of *ensemble consistent explanations*, and we provide a particular realization using counterfactual explanations.

2. We empirically demonstrate the usefulness of ensemble consistent counterfactual explanations in sensor fault detection & localization in water distribution networks.

The remainder of this work is structured as follows: First, we briefly review the necessary foundations in Section 2. Next in Section 3, we propose the abstract concept of *ensemble consistent explanations* and a specific realization thereof using counterfactual explanations, dubbed *ensemble consistent counterfactual explanations*. Following up on this, we empirically evaluate the proposed technology in the real-world task of sensor fault identification in water distribution networks in Section 4, a domain where ensembles arise naturally due to the spatial nature of the task. Finally, we close our work with a summary and conclusion in Section 5.

2 Foundations

As mentioned in the introduction, there exist a wide variety of different explanation methodologies [Molnar, 2019]. A prominent instance of contrasting explanations are Counterfactual explanations (often just called *counterfactuals*), which

---

[1]E.g. looking for anomalies such as leakages and sensor faults that might cause water loss or water contamination.

state a change to some features of a given input such that the resulting data point, called the counterfactual, causes a different behavior of the system than the original input does. Thus, one can think of a counterfactual explanation as a suggestion of actions that change the model's behavior. One reason why counterfactual explanations are so popular is that there exists evidence that explanations used by humans are often contrasting in nature [Byrne, 2019] – i.e. people ask questions like *"What needs to be changed in order to observe a different outcome?"*. In order to keep the explanation (suggested change) simple – i.e. easy to understand – an obvious strategy is to aim for a minimum change so that the resulting sample (counterfactual) is similar/close to the original sample. This is formalized via the following definition Definition 1.

Definition 1 ((Closest) Counterfactual Explanation [Wachter *et al.*, 2017]). *Assume a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ is given (where often $\mathcal{X} = \mathbb{R}^d$). A counterfactual $\vec{x}_{cf} \in \mathcal{X}$ for a given input $\vec{x}_{orig} \in \mathcal{X}$ is given as a solution of the following optimization problem:*

$$\arg\min_{\vec{x}_{cf} \in \mathcal{X}} \ell\left(h(\vec{x}_{cf}), y_{cf}\right) + C \cdot \theta(\vec{x}_{cf}, \vec{x}_{orig}) \quad (1)$$

where $\ell(\cdot)$ denotes a loss function, y_{cf} the target prediction, $\theta(\cdot)$ a penalty for dissimilarity of $\vec{x}_{cf}$ and $\vec{x}_{orig}$, and $C > 0$ denotes a regularization strength.

Such counterfactuals are also called *closest counterfactuals* [Karimi *et al.*, 2021; Artelt and Hammer, 2019]. Despite their popularity, counterfactual explanations have some weaknesses. One is posed by a missing uniqueness: There can exist more than one valid counterfactual explanation – this is called Rashomon effect [Molnar, 2019]. In such cases, it is not clear which or how many of those should be presented to the user. One common modeling approach is to ignore this effect, or to add a suitable regularization. If an ensemble of decisions should be explained, the Rashomon effect can yield inconsistent individual explanations, thus constitutes a severe problem. Therefore, we propose a modeling which aims for consistency of such explanations.

3 Ensemble Consistent Explanations

In the following, we assume that we are given a set (i.e. an ensemble) of decision functions $h_i \in \mathcal{H}$, $h_i : \mathcal{X} \rightarrow \mathcal{Y}$ together with a sample $\vec{x} \in \mathcal{X}$ – as an example, this might be an ensemble in space or time. We are looking for a single local explanation that consistently explains the behavior of all decision functions $h_i(\cdot)$ for a given input state $\vec{x}$.

We assume $\Gamma(\vec{x})$ refers to the set of all possible local explanation at a given input state $\vec{x} \in \mathcal{X}$, which can be derived from any of the decisions within the ensemble. There exist two contrasting objectives, to choose a "consistent optimum" from this set: (i) how complex is an explanation (e.g. as measured by the distance to the original input state); we assume that $\theta : \Gamma \rightarrow \mathbb{R}_+$ is a function that measures the complexity of a given explanation; smaller values correspond to a "low complexity" explanation while the opposite is true for "high complexity" explanation. (ii) What is the fidelity of an explanation w.r.t. all decisions within the ensemble, i.e. does it explain the behavior of every given function $h_i(\cdot)$ at a

given sample or not. We assume that $\Psi : \Gamma \times \mathcal{H} \times \mathcal{X} \rightarrow \mathbb{R}$ is a function that measures "how well" the explanation explains the behavior of a given function $h_i(\cdot)$ at a given sample. We assume that $\Psi(\cdot) \geq 0$ holds if the explanation is "correct" (e.g. the computed value constitutes a counterfactual) and $\Psi(\cdot) < 0$ otherwise.

An *ensemble-consistent explanation* is given by a single, low complexity, local explanation $\gamma \in \Gamma(\vec{x})$ that explains the behavior/prediction of all given decision functions $h_i(\cdot)$, i.e.

$$\min_{\gamma \in \Gamma(\vec{x})} \theta(\gamma)$$
$$\text{s.t. } \Psi(\gamma, h_i, \vec{x}) \geq 0 \quad \forall i \quad (2)$$

In practice, this problem needs to be relaxed to guarantee at least one feasible solutions. Therefore we use slack variables ξ_i that allow constraint violations, which are weighted by a penalty $\lambda > 0$:

$$\min_{\gamma \in \Gamma(\vec{x})} \theta(\gamma) + \lambda \sum_i \xi_i$$
$$\text{s.t. } \Psi(\gamma, h_i, \vec{x}) \geq 0 - \xi_i \quad \forall i \quad (3)$$
$$\xi_i \geq 0 \quad \forall i$$

3.1 Consistent Counterfactual Explanations

In the following, we use counterfactual explanations (see Definition 1 in Section 2) as a specific type of local explanations for realizing the idea of ensemble consistent explanations as described in the previous section.

Because we consider a set of prediction functions $h(\cdot)_i$, we might request potentially different target predictions y_{cf_i} in the counterfactual explanation.

Classification
Assuming a set of binary classifiers $h_i : \mathbb{R}^d \rightarrow \{-1, 1\}$. We can implement the idea of ensemble consistent explanations by refining Eq. (3) as follows:

$$\min_{\vec{x}_{cf} \in \mathbb{R}^d} \theta(\vec{x}_{cf}, \vec{x}) + \lambda \sum_i \xi_i$$
$$\text{s.t. } h_i(\vec{x}_{cf}) \cdot y_{cf_i} \geq 0 - \xi_i \quad \forall i \quad (4)$$
$$\xi_i \geq 0 \quad \forall i$$

Regression
Similar, for a set of regression functions $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$, we refine Eq. (3) as follows:

$$\min_{\vec{x}_{cf} \in \mathbb{R}^d} \theta(\vec{x}_{cf}, \vec{x}) + \lambda \sum_i \xi_i$$
$$\text{s.t. } d\left(f_i(\vec{x}_{cf}) - y_{cf_i}\right) \leq \delta_i + \xi_i \quad \forall i \quad (5)$$
$$\xi_i \geq 0 \quad \forall i$$

where $d : \mathbb{R} \rightarrow \mathbb{R}_+$ denotes a function for penalizing prediction errors – e.g. the absolute error or the squared error[2] – and $\delta_i \geq 0$ denotes a tolerance threshold on the prediction error of the i-th regression function, since it might be numerically difficult to achieve exactly y_{cf_i} as a prediction.
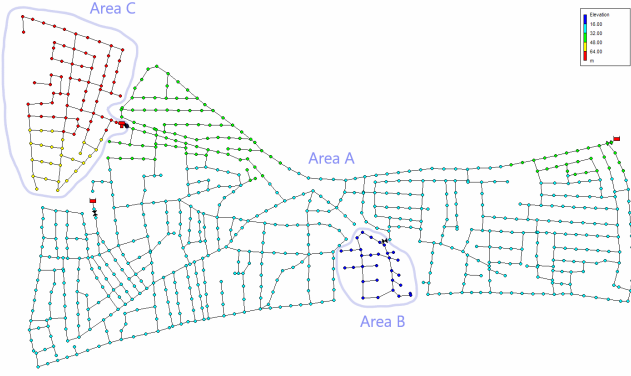
---

[2]$d(a) = a^2$ or $d(a) = |a|$

Figure 1: L-Town network [Vrachimis *et al.*, 2020] – we only use the hydraulically isolated "Area A", where 29 pressure and 2 flow sensors are installed.
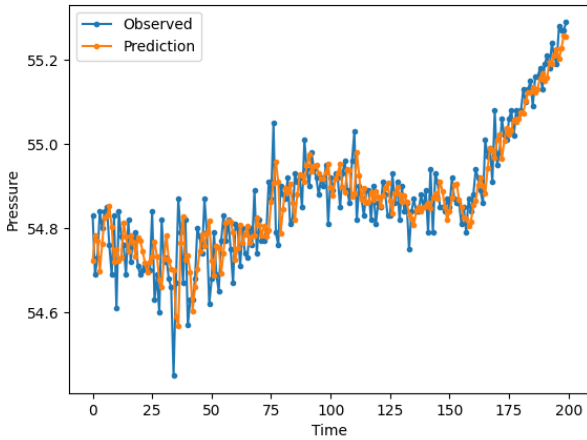

Figure 2: Illustration of the a virtual pressure sensor: Predicted vs. observed pressure values at node *n549*.

4 Experiments

We empirically evaluate our proposed *ensemble consistent counterfactual explanations* in the important real world application of sensor fault detection and localization in water distribution networks (WDNs) – in this context, we use our proposed explanations for identifying a faulty sensor. The implementation of the experiments is publicly available[3].

4.1 Setup

As a WDN, we use a variation of the popular L-Town water distribution network [Vrachimis *et al.*, 2020] – for simplicity, we demonstrate our approach on the hydraulically isolated Area A, by closing the valves that provide water to Area B and C (see Fig. 1). Area A consists of 661 nodes and 766 links. The benchmark assumes the existence

---

Metric	Score (avg. $\pm$ variance)
True Positives	1.0000 ± 0.0000
True Negatives	0.9951 ± 0.0009
False Positives	0.0000 ± 0.0000
False Negatives	0.0049 ± 0.0009
Detection delay	1.0021 ± 0.0020

Table 1: Evaluation of anomaly detection

of 29 pressure sensors at nodes (optimally placed to maximize the detectability of leakages) and 2 flow sensors at the water inlets of the network. We implement a residual based anomaly detection method by means of virtual sensors [Eliades and Polycarpou, 2012; Santos-Ruiz *et al.*, 2019; Vaquet *et al.*,]. We build a virtual sensor $f_i : \mathbb{R}^{30} \to \mathbb{R}$ for each of the 29 pressure sensors – the virtual sensor tries to predict the pressure at its specific location based on the previous measurements from all other 28 (i.e. $29 - 1$) pressure and 2 flow sensors (see Fig. 2). The virtual sensors are implemented using linear regression, we use a sliding window of size $T = 3$ time steps for averaging the other past sensor values which are then fed into the linear regression:

$$ f_i(\vec{x}_{t-1}^i) = \vec{w}_i^\top \vec{x}_{t-1}^i + b_i \text{ with } \vec{x}_{t-1}^i = \frac{1}{T}\sum_{j=1}^{T}(\vec{x}_{t-j})_{\neq i} \quad (6) $$

where $(\vec{x}_{t-j})_{\neq i}$ denotes the sensor measurements at time $t - j$ at all nodes except the i-th node. The anomaly detection itself is realized using a simple threshold δ that is estimated during training on a fault free time window – we raise an alarm whenever any predicted pressure $f_i(\vec{x}_{t-1}^i)$ (from the virtual sensor) deviates from the observed pressure $(\vec{y}_t)_i$:

$$ \|f_i(\vec{x}_{t-1}^i) - (\vec{y}_t)_i\|_\infty > \delta \quad (7) $$

We generate 434 scenarios using the variation of the L-Town network where we place sensor faults of different magnitude on the pressure sensors – we consider three different magnitudes for each of the following five types of sensor faults [Reppa *et al.*, 2016]:

1. A constant offset of the sensors measurement compared to the true measured quantity.

2. Gaussian noise added to the sensor measurements.

3. Sensor power failures, which results to the measurement being equal to zero.

4. An offset of the sensor measurement, linearly proportional to the true measured quantity.

5. Sensor drift, resulting in an increasing measurement value until a maximum limit is reached.

Each scenario is exactly three month long and contains exactly one sensor fault – the first month is always fault free, so it can be used for training the virtual sensors.

4.2 Results

First, we evaluate the performance of the implemented residual based anomaly detection by reporting the mean and variance of the confusion matrix and detection delay (time until
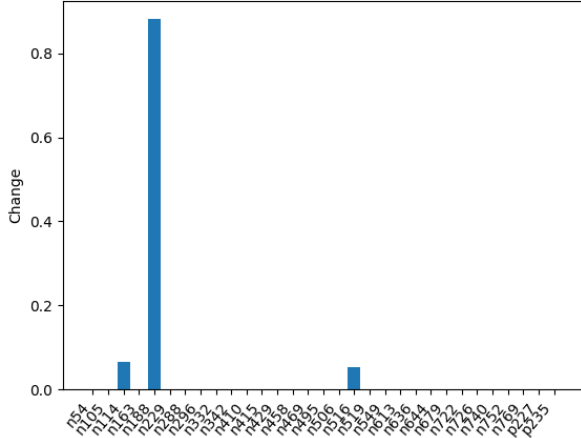
Figure 3: Distribution of predicted faulty sensors according to the independent counterfactuals for a raised alarm in case of a sensor fault at node *n332* – note that the predictions are wrong.

Metric	Score (avg. $\pm$ variance)
Our method	0.97 ± 0.03
Baseline	0.00 ± 0.00

Table 2: Evaluation (in %) of successful sensor fault localization.

the first alarm is raised) over all scenarios – the results are shown in Table 1. We observe that our implemented anomaly detector is consistently able to detect all sensor faults with an average delay of 1-2 time steps. Next, we evaluate the fidelity of our proposed *ensemble consistent counterfactual explanations* by trying to identify the faulty sensor from the explanation – i.e. identifying the cause of the observed anomaly. For every detected fault (i.e. raised alarm), we compute an ensemble consistent counterfactual explanation by solving Eq. (5) using convex programming and normalize it for convenience – see Fig. 4 for an example. That is we are looking for a single explanation that tells us how to change the sensor measurements such that no alarm would have been raised. Note that since an anomaly stays active for some time, we have a sequence of permanent alarms. We select the sensor with the largest magnitude of proposed changes (according to the explanation) as our prediction of the faulty sensor which we then compare with the known ground truth. Data and predictions are a bit noisy (see Fig. 2), hence other sensor components might be non-vanishing as well, as can be seen in Fig. 4. We compare the consistent explanation to the baseline which is obtained by aggregating independent local counterfactual explanations for every decision: that means, we compute a counterfactual explanations of each model separately, and select the sensor with the largest change as an estimate of the faulty sensor – we aggregate these estimates by selecting the most common estimate of the faulty sensor. The results are shown in Table 2. We observe that, in contrast to the baseline which is not capable of identifying the faulty sensor, we are able to correctly identify the faulty sensor in almost
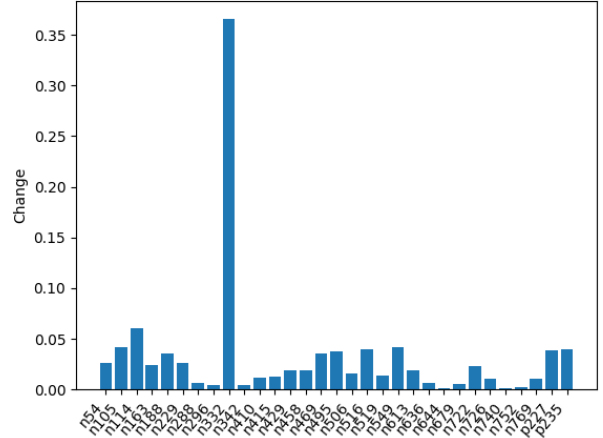

Figure 4: Illustration of a normalized *ensemble consistent counterfactual explanations* for a raised alarm in case of a sensor fault at node *n332*.

every scenario. This fact demonstrates the high fidelity and usefulness of our proposed explanation methodology in this application example. It turns out that the independent counterfactuals are very sparse but almost always have the peak at the same (wrong) sensor (see Fig. 3 for an example) – i.e. they exploit some model specific properties to get a sparse counterfactual, which is unrelated to the faulty sensor.

5 Summary & Conclusion

In this work we proposed the concept of *ensemble consistent explanations* for computing explanations of an ensemble of decisions instead of a single decision only. Ensembles naturally occur for processes with spacial distribution – i.e. a set of models such as virtual sensors in networks as we considered in our experiments – or in time – i.e. different instances of a single model evolving over time. Furthermore, we also proposed a specific implementation of this concept using counterfactual explanations yielding *ensemble consistent counterfactual explanations*. For the purpose of an empirical evaluation, we applied our proposed explanations to the (important real-world) problem of sensor fault detection and localization in water distribution networks. We observed that our proposed explanations show very good performance for explaining the detected anomalies and identifying the faulty sensors which caused the observed anomaly.

A couple of extensions and directions of further research can be based on this initial work. In particular, driven by the specific application domain, we aim for the following: How does the method perform for more than one sensor fault, i.e. explanations which need to identify multiple features or more complex signals? How is the behavior for more complex nonlinear dynamics modeled by deep networks? What are suitable extensions to temporal dynamics and ensembles in time rather than space, such as tackled using deep recurrent models or attention mechanisms? Are there formal guarantees to substantiate consistency of the proposed method?

Acknowledgments

References

[Aamodt and Plaza., 1994] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and systemapproaches. *AI communications*, 1994.

[Artelt and Hammer, 2019] André Artelt and Barbara Hammer. On the computation of counterfactual explanations - A survey. *CoRR*, abs/1911.07749, 2019.

[Byrne, 2019] Ruth M. J. Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI-19*, 2019.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[Eliades and Polycarpou, 2012] DG Eliades and Marios M Polycarpou. Leakage fault detection in district metered areas of water distribution systems. *Journal of Hydroinformatics*, 14(4):992–1005, 2012.

[Fisher *et al.*, 2018] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. *arXiv e-prints*, page arXiv:1801.01489, Jan 2018.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), August 2018.

[Karimi *et al.*, 2021] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)*, 2021.

[Molnar, 2019] Christoph Molnar. *Interpretable Machine Learning*. 2019.

[Offert, 2017] Fabian Offert. "i know it when i see it". visualization and intuitive interpretability, 2017.

[parliament and council, 2016] European parliament and council. General data protection regulation: Regulation (eu) 2016/679 of the european parliament, 2016.

[Perez *et al.*, 2014] Ramon Perez, Gerard Sanz, Vicenc Puig, Joseba Quevedo, Miquel Angel Cuguero Escofet, Fatiha Nejjari, Jordi Meseguer, Gabriela Cembrano, Josep M. Mirats Tur, and Ramon Sarrate. Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona [applications of control]. *IEEE Control Systems Magazine*, 34(4):24–36, 2014.

[Reppa *et al.*, 2016] Vasso Reppa, Marios Polycarpou, and Christos Panayiotou. Sensor fault diagnosis. *Foundations and Trends in Systems and Control*, 3:1–248, 01 2016.

[Samek *et al.*, 2017] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017.

[Santos-Ruiz *et al.*, 2019] Ildeberto Santos-Ruiz, Francisco-Ronay López-Estrada, Vicenç Puig, and Joaquim Blesa. Estimation of node pressures in water distribution networks by gaussian process regression. In *2019 4th Conference on Control and Fault Tolerant Systems (SysTol)*, pages 50–55. IEEE, 2019.

[Sanz *et al.*, 2016] Gerard Sanz, Ramon Pérez, Zoran Kapelan, and Dragan Savic. Leak detection and localization through demand components calibration. *Journal of Water Resources Planning and Management*, 142(2):04015057, 2016.

[Vaquet *et al.*,] Valerie Vaquet, André Artelt, Johannes Brinkrof, and Barbara Hammer. Taking care of our drinking water: Dealing with sensor faults in water distribution networks.

[Verma *et al.*, 2020] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review, 2020.

[Vrachimis *et al.*, 2020] Stelios G Vrachimis, Demetrios G Eliades, Riccardo Taormina, Avi Ostfeld, Zoran Kapelan, Shuming Liu, Marios Kyriakou, Pavlos Pavlou, Mengning Qiu, and Marios M Polycarpou. Battledim: Battle of the leakage detection and isolation methods. 2020.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.