Learning Relevant Explanations

Chris Russell [1] **Rory Mc Grath** [2] **Luca Costabello** [2]

Abstract

We present a novel approach for learning to generate counterfactual explanations of decisions made by classifiers. We show how to improve user scoring of explanations by using ranking data over pairs of diverse explanations. Applying this framework to a range of classification problems, we show how varying explanations can be learnt depending on their intended application. We empirically demonstrate the importance of context and that the precise form of a good explanation depends upon what it is to be used for.

Counterfactual Explanations are a recent approach to explaining black-box classifiers. Such approaches explain why an algorithm has made a particular decision, by showing one or multiple ways in which the data could be altered to receive an alternate decision. These Counterfactual Explanations make use of Lewis (1973) "Closest Possible world" formulation. That is, as an explanation for why a particular decision was made, they offer the closest or most similar datapoint for which another decision was made. For example, a counterfactual explanation of why a black-box algorithm denied a loan could take the form Wachter et al. (2018):

> You were denied a loan because you have an income of $30,000. If you had an income of $45,000 you would have been offered the loan.

These explanations have the potential to provide direct and comprehensible explanations about the decisions made by algorithms without requiring the person receiving the explanation to understand the internal logic of the decision making system. For most complex systems, there are multiple ways to obtain an alternative outcome, and consequentially multiple counterfactuals could be offered as competing explanations. Lewis's (1973) notion of selecting the "closest possible world" as the counterfactual offers one way to resolve this. However, this simply replaces the ambiguity of which world to select as a counterfactual, with the ambiguity of what should we use as a measure of closeness. Lewis has received criticism both from philosophy and the machine-learning communities for this limitation. Pearl (2000) offers this criticism as a justification for preferring Structured Equation Models for causal reasoning over Lewis's notion:

> "However, the closest-world semantics still leaves two questions unanswered. (1) What choice of distance measure would make counterfactual reasoning compatible with ordinary conceptions of cause and effect? (2) What mental representation of interworld distances would render the computation of counterfactuals manageable and practical (for both humans and machines)?"

We directly address the limitations of the "closest possible world" model of counterfactuals raised by Pearl, in the context of counterfactual explanations. We show how a distance measure can be learnt from user preferences to offer relevant explanations. Finally, we show the importance of context for automatically generated explanations, and that the choice of distance measure depends upon its intended use.

1. Prior Work

We follow Lewis (1973) and formulate a counterfactual[1] as a "close possible world" in which a different classifier response occurs. As is common (Wachter et al., 2018; Russell, 2019; Ustun et al., 2019) we write this as: Given a datapoint x, the closest counterfactual c can be found by solving:

$$\arg\min_{c} d(x, c) \quad \text{such that: } f(c) = r \tag{1}$$

Here $d(\cdot, \cdot)$ is a distance measure, typically a weighted ℓ_1 norm[2], f the classifier function we are computing a coun-

[1]Amazon Tübingen. Some of this work was done at the Alan Turing Institute and the University of Surrey [2]Accenture. Correspondence to: Chris Russell <cmruss@amazon.de>.

2020 ICML Workshop on Human Interpretability in Machine Learning (WHI 2020).

[1]For a philosophical treatment of counterfactuals as explanation, see Kment (2006).

[2]Wachter et al. (2018) proposed the widely used heuristic of weighting components of the ℓ_1 norm by the inverse median absolute deviation of the variable computed over the dataset. This measure is invariant to linear transforms of individual variables and robust to outliers. See discussion in Russell (2019) for adapting it to binary variables where the MAD is guaranteed to be 0.

terfactual for, and r the response we desire the classifier function $f(\cdot)$ to take in this closest possible world.

Using counterfactual explanations to explain the decisions made by black-box classifiers has attracted attention outside of machine learning. Wachter et al. (2018) being cited within the official Guidelines on Automated individual decision-making and Profiling in the GDPR (Various, 2018a); while the Google What-if tool (Various, 2018b) offers a variant on counterfactual explanations as a tool to let non-experts better understand machine learning models.

Binns et al. (2018) and Lim & Dey (2009) performed user studies of explanations with both finding a preference for counterfactual explanations[3]. This preference for counterfactual explanations is perhaps unsurprising. In their review of folk, i.e. everyday, explanations of behaviour, Malle (2011) found that 80% of explanations are concerned with the intentionality of actors and the remaining 20% are forms of causal history reasons, a type of explanation related to counterfactual explanations. With notions of intentionality difficult to apply to most machine learning, counterfactuals and other forms of causal history reason may be the only form of folk explanation directly applicable to understanding the behaviour of algorithms. No prior work has used user preferences to improve generated explanations.

The importance of context in everyday folk explanation has been well addressed in the philosophical (Grice et al., 1975; Hilton, 1990) and psychological (Malle, 2011) literature, and for a good overview of some issues from a machine learning perspective see Miller (2017). However, while some works (Weller, 2017; Ribera & Lapedriza, 2019) give potential uses for explanations, there has been no systematic evaluation showing which context (or intended use) particular forms of explanation best suit, and no attempt to learn more relevant forms of explanation for a given use.

2. Finding Relevant Counterfactuals

One challenge in learning to offer relevant explanations is the unavailability of ground-truth data. Given a datapoint and classifier response, people typically cannot describe what a relevant explanation should look like. However, given a set of explanations, and an intended use case, such as: "make me trust the system more"; "make me more aware of ethical issues within a system"; or "making it easier for people to make a change that would get them a different outcome in the future", people can select what they believe is the most useful explanation. This leads us naturally to consider human-guided ranking formulations, in which we learn a penalty function C so that the most useful explanation found c has a lower penalty associated with it. We combine this with methods for generating low-penalty counterfactuals on the fly and iteratively train by reranking these new candidate explanations.

[3]Such explanations are called "sensitivity" by Binns et al. (2018) and "why-not explanations" by Lim & Dey (2009).

Generalised Counterfactual Penalties Replacing hand-crafted penalty measures with learnt objectives allows us to consider a broader family of possible penalty functions than is usual.Given x an input datapoint, c the counterfactual to be found for datapoint x, and learnt parameters Θ, we define a penalty function $C_\Theta(c, x)$ as:

$$C_\Theta(c, x) = d_\Theta(c, x) + F_\Theta(c) \tag{2}$$

$C(c, x)$ generalises the penalty of previous works by: *(i)* relaxing the symmetry constraints of true distance measures and allowing the pseudo-distance $d(c, x)$ to be asymmetric; *(ii)* it includes a *Fidelity term* $F(c)$ which encourages the counterfactual to take particular values.

We use an asymmetrically weighted variant of the ℓ_1 distance to encourage the sparsity of the found counterfactual.

$$d_\Theta(x, c) = w_1 \cdot \text{reLu}(x - c) + w_2 \cdot \text{reLu}(c - x) \tag{3}$$

where $w_1, w_2 \in \Theta$ are learnt weight vectors.

The fidelity term $F(c)$ encourages the found counterfactual to be more representative of the underlying data while preserving a preference for sparsity. $F(c)$ is defined as:

$$F_\Theta(c) = w_1' \cdot \text{reLu}(m - c) + w_2' \cdot \text{reLu}(c - m) \tag{4}$$

where $m \in \Theta$ is a learnt vector that encourages c to take values close to m, with the strength of this attraction depending on weights w_1' and w_2'.

Learning Relevant Counterfactuals To generate relevant counterfactuals, we adopt a ranking formulation in which we learn the learnt parameters of the penalty function $C_\Theta(c, x)$ so that the most useful explanation found c has a lower penalty associated.

We define a preference tuple $t = (x, c^{(1)}, c^{(2)})$ as a datapoint x, and two counterfactuals $c^{(1)}$ and $c^{(2)}$ such that $c^{(1)}$ is preferred over $c^{(2)}$ (we collect such preferences from humans, by means of user surveys, Section 3). We define an empiric loss over the set $\mathcal{T}$ of all preference tuples t as:

$$P(i) = \begin{cases} 1 & \text{if user preferred } c_i^{(1)} \text{ over } c_i^{(2)} \\ -1 & \text{otherwise.} \end{cases} \tag{5}$$

$$\mathcal{L}_\Theta = \sum_{(x, c_i, c_j) \in \mathcal{T}} \text{reLu}(1 + P_i C_\theta(c_i^{(1)}, x_i) - P_i C_\Theta(c_i^{(2)}, x_i)) \tag{6}$$

This is the paired hinge loss used in a ranking SVM (Joachims, 2002) adapted for our distance measure $C_x(c)$.

This loss encodes a soft penalty that says the cost of the preferred explanation must be 1 less than the cost of the other explanation, otherwise you pay a penalty linear in the size of the violation. We regularise the coefficients with an ℓ_2^2 penalty, slightly modified to encourage non-zero weights in the vectors w_1 and w_2:

$$\mathcal{R}_\Theta = \rho \left(\sum_k (w_{1,k} - 0.1)^2 + \sum_k (w_{2,k} - 0.1)^2 \right) + \sum_k {w'}^2_{1,k} + \sum_k {w'}^2_{2,k} + \sum_k m_k^2 \quad (7)$$

where ρ is a small scalar coefficient we set to 0.01.

Counterfactuals Generation The above by itself is insufficient to guarantee that the minimal solution found will be useful, as the optimiser could find a new form of explanation that does not exist in the training set $\mathcal{T}$. As such, we leverage human preferences to iteratively update the training set with newly generated counterfactual explanations. We use the learnt weights Θ to generate a new relevant counterfactual c for each tuple $(x, c^{(1)}, c^{(2)}) \in \mathcal{T}$, by minimising (2) under the constraint that the model must give the desired outcome. Newly generated counterfactuals are annotated to create additional preference tuples $(x, c^{(1)}, c^{(2)})$ that we append to the training set $\mathcal{T}$. We proceed iteratively, and stop on convergence or after a preset number of iterations. Pseudo-code is available in Algorithm 1 in the Appendix.

3. Experiments

We focus on three tasks: the first being what Wachter et al. (2018) described as "to understand what could be changed to receive a desired result in the future," and Ustun et al. (2019) called "actionable recourse"; the second being to flag inactionable factors in the decision-making system, such as an explicit dependency on the race; age; or sexuality of the subject; and finally to increase the users' trust in the system.

There is much literature on algorithmic fairness (Narayanan, 2018), and the notion of fairness is more complex than simply checking if an algorithm exhibits a dependency on an attribute such as race. Nonetheless, these counterfactual explanations can help with understanding the behaviour of algorithms with respect to race, and in providing concrete examples to a lay audience that might have greater impact than statistics. As such, some of the examples in this section may be offensive. We report them frankly, as it makes explicit the behaviour of standard algorithms trained on biased data, and, moreover, shows how this behaviour can be masked by various choices of metric.

Illustrative Example We show results on the relatively simple and strongly racially-biased LSAT dataset (Bock & Lieberman, 1970). This historic dataset consists of three variables: the score in a law-school entrance exam (LSAT); grade point average (GPA); and whether or not a student identified as being black. The task is to predict whether or not a student's first-year average score will be better or worse than average. The dataset is noticeable for having such a strong racial bias that logistic regression predicts that any student who identifies as black will perform worse than average regardless of their GPA or LSAT.

You were predicted to do better than average. Some ways you could have been predicted to do worse were:

Explanation Type	Closest Explanation	Second Closest	Furthest Explanation
Actionable	LSAT took value 31 rather than 45	GPA took value 1.6 rather than 3.3	if you were black
Ethically Concerning / Inactionable	if you were black	LSAT took value 31 rather than 45	GPA took value 1.6 rather than 3.3
Trustworthy	GPA took value 1.6 rather than 3.3	LSAT took value 31 rather than 45	if you were black

Table 1: Using our method to generate diverse counterfactuals over a single individual on the LSAT dataset. All explanations are accurate and describe different dependencies of the same decision.

We hand-labelled twenty pairs of counterfactuals, generated using a variant of Russell (2019): if the counterfactuals felt actionable; demonstrated a racial bias; or made the classifier seem trustworthy. Table 1 shows the full set of counterfactuals generated for a positively classified individual.

Although the total set of counterfactuals found by each method remains unchanged if only a single counterfactual was offered as an explanation, each choice of measure would present a very different view of the behaviour of the algorithm, with two of the three measures masking a strong explicit racial bias in the classifier[4].

Learning from User Preferences We adopt three publicly-available tabular datasets widely used in the fairness and explainability literature: IBM HR Analytics Employee Attrition & Performance Data; Graduate School Admission; COMPAS Recidivism. For each dataset a binary classifier is trained using logistic regression to predict if an employee is likely to remain in the company; if a student is likely to be admitted to grad school; or if an individual is likely to be placed in the high-risk group of reoffending.

[4]Owing to the aforementioned strong racial bias in this dataset, no masking of the racial bias is possible for black students. The counterfactual would necessarily include a change of race as a prerequisite for being predicted to do better than average by the classifier.

	Actionability	Trust	Inactionable
Admissions	$88\% \pm 4.6\%$	$72\% \pm 6.3\%$	$74\% \pm 6.2\%$
COMPAS	$72\% \pm 6.3\%$	$96\% \pm 2.8\%$	$68\% \pm 6.6\%$
IBM HR	$80\% \pm 5.7\%$	$73\% \pm 6.4\%$	$56\% \pm 7.0\%$

Table 2: Human validation: (Russell, 2019) vs relevant explanations. The rate at which annotators preferred our counterfactual explanation over the baseline. Each case was scored with 50 annotations except 'IBM,Trust' where annotators were unsure in 2 cases, with values computed over the remaining 48.

We learn three distinct relevant counterfactual generators:

1. **Actionability**: Which explanation would be the most helpful in telling you what you could change to get a new outcome in the future?
2. **Trust**: Which explanation makes you trust the system the most?
3. **Inactionable**: Which explanation makes you least hopeful that an alternative outcome is possible?

We initialize each training set of preference tuples $\mathcal{T}$ by sampling from a held-out set of 50 candidate points x that received an automated decision (i.e. 'Your admission was rejected'). We create three initial unique counterfactuals for each candidate datapoint x.

At each iteration, we ask 5 human annotators to complete a preference survey where they are presented a hypothetical scenario x and pairs of counterfactuals to rank. The baseline counterfactual is generated using Equation 1, with a ℓ_1 distance measure as described in footnote 2. Each of the chosen 5 participants is asked to evaluate 10 of such pairs of counterfactuals, for a total of 50 evaluated pairs for each iteration. Participants can skip a pair if they feel unsure about their choice. Each participant is asked to spend a minimum of 30 seconds per pair before recording their decision. Annotators are randomly chosen from a pool of 10 members of a research institution. 7 among them are research engineers, 2 scientists, and 1 workshop organiser. Participants are 50% female, and belong to 8 different nationalities.

Results We measure how often annotators choose the learnt counterfactual option over the baseline. Table 2 reports accuracy results for the final iteration, on each dataset/criterion combination. Although the preference-aware counterfactuals are always more desirable, we also noted that in a few cases (particularly 'inactionable' for IBM HR) they barely outperform the baseline. There are two obvious possibilities here, the first is that the baseline method may offer good explanations by default, while a second possibility is that there are situations where there are multiple good choices due to the features used. An example of this can be seen with COMPAS, where counterfactuals trained to identify inactionable explanations vary both age and ethnicity: annotators are unable to reach a consensus if a decision due to age is more inactionable than a decision due ethnicity. What is particularly striking about results on COMPAS, is that there seems to be no actionable answers (supplementary materials Sec. C), and although there seems to be a uniform preference for age being varied in all types of explanation, perhaps because it is easy to comprehend and less offensive than alternatives such as varying race, there is little difference in what is learnt for actionable, trustworthy, or inactionable answers (Table 4). Despite the absence of unique compelling explanations for each context, there are clearly bad explanations, with our explanation being found more trustworthy than Russell (2019) 96% of the time.

Table 3 provides examples of how preference-aware counterfactuals evolved from the initial iteration to the final one. Table 4 shows the preferred features learned for each dataset, i.e. the set of features used to generate preference-aware counterfactuals for the desired preference criterion. Our method successfully learns to promote diverse feature sets for each criterion. We note that in Graduate Admission, a student's cumulative grade point average (CGPA) from their studies was frequently varied in explanations both designed to promote trust and those that make the system's decision appear inalterable. With the benefit of hindsight, this is perhaps unsurprising: providing actionable explanations that emphasise the ease of changing the decisions made by the system can make such system appear arbitrary, easy to game, and consequentially less trustworthy.

4. Conclusion

We have presented a novel approach for learning relevant explanations. As autonomous systems become more ubiquitous there has been an increasing push towards putting "humans in the loop" and using people to validate decisions made by these systems. By choosing explanations to help people contest decisions, and by, for example, flagging dependencies on data that is likely to be incorrect, we can empower people to contest the results of algorithms. At the same time our results are concerning: on the COMPAS dataset, which describes the behaviour of a hotly-contested and racially biased automated system, we can improve reported trust in a näive classifier in 96% of cases without altering the behaviour of the classifier. As work in this field moves forward it becomes more apparent we should not just call for explanations of algorithms as a tool to increase accountability, but also ask why particular explanations are being offered.

This work was partially supported by the Omidya Group and The Alan Turing Institute under the EPSRC grant no. EP/N510129/1

	Initial Iteration	Final Iteration
Actionability	Job Involvement had been 3.5/5 rather than 2/5.	Overtime had been No rather than Yes.
Inactionable	Business Travel had been 'No Travel' rather than 'Frequently Travel'.	Job Involvement had been 4/5 rather than 2/5; Job Satisfaction had been 2.7/5 rather than 1/5.
Trust	Years In Current Role had been 8 rather than 0.	Job Involvement had been 4/5 rather than 3/5; Num Companies Worked had been 0 rather than 1; Work Life Balance had been 3.4/5 rather than 3/5.

Table 3: Example pairs from initial and final iterations, IBM HR Dataset. Relevant explanations better reflect the desired preference criterion. Examples for other datasets are reported in Appendix C.

	Actionability	Trust	Inactionable
IBM HR	Environment Satisfaction, Over Time, Business Travel	Distance From Home, Job Role, Relationship Satisfaction, Training Times Last Year	Job Involvement, Job Satisfaction
Graduate Admission	GRE, TOEFL	CGPA	CGPA
COMPAS	Age	Age	Age, Ethnicity

Table 4: The features most commonly selected by our method when generating preference-aware relevant counterfactuals.

Supplementary Material for 'Learning Relevant Explanations'

A. Criteria for labelling LSAT explanations

Pairs of explanation were labelled according to the following criteria:

1. Counterfactuals that only varied entry exam results were assumed to be most actionable; while those that varied race were assumed to be least actionable.
2. Counterfactuals that varied race were assumed to be most indicative of a racial bias, otherwise a coin was flipped to select a counterfactual.
3. Counterfactuals that varied race were assumed to make people least likely to trust the classifier; otherwise, a coin was flipped.

B. Questions Asked to Human Annotators

Table 5 reports the questions asked to human annotators for each dataset/criteria combination.

C. Initial vs Final Iteration: Graduate Admission, COMPAS

We report additional examples of preference-aware explanations generated on Graduate Admission and COMPAS. Results for IBM HR are reported in Table 3 in the main paper.

Algorithm 1: Preference-aware Counterfactual Generation

Data: An initial solution set of factuals X, and their labels Y, both indexed by i.
Result: Optimal parameters Θ
def $D(i)$**:**
  The diverse set of counterfactuals for point x_i (Section 3)
def $M_\Theta(i)$**:**

$$\arg\min_c C_\Theta(c, x_i) \text{ such that } f(c) = 1 - y_i$$

$$\mathcal{T} := \{(c_i^{(1)}, c_i^{(2)}) : \ c_i^{(1)}, c_i^{(2)} \text{ are sampled from } D(i) \forall x_i \in X\}$$

$$P(i) := \begin{cases} 1 & \text{if user preferred } c_i^{(1)} \text{ over } c_i^{(2)} \\ -1 & \text{otherwise.} \end{cases}$$

foreach *iteration* **do**

$$\Theta := \arg\min_\Theta \mathcal{L}_\Theta + \mathcal{R}_\Theta$$

$$\mathcal{T}' := \{(M_\Theta(i), c_i) : \ c_i \text{ is sampled from } D(i) : \ \forall x_i \in X\}$$

$$P'(i) := \begin{cases} 1 & \text{if user preferred } M_\Theta(_i) \text{ over } c_i \\ -1 & \text{otherwise.} \end{cases}$$

$$\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{T}'$$

$$P \leftarrow P \cup P'$$

end

	Actionability	Trust	Inactionable
IBM HR	Which option do you think you could more easily implement?	Which option do you think is a better offer for the employee?	Which option do you think is more ethically acceptable for the employee?
Graduate Admission	Which option do you think you could achieve for your application next year?	Which option do you think is a better reason for your application being rejected?	Which option makes you think you will never get accepted into the Masters Programme?
COMPAS	Which option do you think you could achieve for your next parole hearing in two years?	Which option do you think is a better reason for being denied parole?	Which option makes you think you are never getting out of prison?

Table 5: We instantiated preference criteria listed in Section 2 as questions tailored to dataset-specific scenarios, to enhance the comprehension of the annotation task. We asked human annotators to impersonate an HR representative (IBM HR), a rejected student (Graduate Admission), an inmate whose parole has been rejected (COMPAS).

	Initial Iteration	Final Iteration
Actionability	Your Ethnicity had been Caucasian rather than African-American.	You were 8 years older.
Inactionable	Your Custody Status had been Probation rather than Jail Inmate.	You were 1 year older.
Trust	Your Ethnicity had been Caucasian rather than African-American.	You were 10 years older.

	Initial Iteration	Final Iteration
Actionability	Your Cumulative Grade Point Average (CGPA) for your Undergraduate Degree had been 10/10 rather than 8/10.	You performed 6% better in your General Test (GRE); You performed 13% better in your Test of English as a Foreign Language (TOEFL); Your Statement of Purpose had scored a 5/5 rather than 2/5.
Inactionable	The University you attended for your Undergraduate Degree had a rating of 3.8/5 rather than 3.0/5.	Your Cumulative Grade Point Average (CGPA) for your Undergraduate Degree had been 9/10 rather than 8/10.
Trust	You performed 4% better in your Test of English as a Foreign Language (TOEFL).	Your Cumulative Grade Point Average (CGPA) for your Undergraduate Degree had been 9/10 rather than 8/10.

Table 6: Initial vs final iteration: COMPAS (top), Graduate Admission (bottom)

References

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 377. ACM, 2018.

Bock, R. D. and Lieberman, M. Fitting a response model forn dichotomously scored items. *Psychometrika*, 35(2): 179–197, 1970.

Grice, H. P., Cole, P., Morgan, J. L., et al. Logic and conversation. *1975*, pp. 41–58, 1975.

Hilton, D. J. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65, 1990.

Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142. ACM, 2002.

Kment, B. Counterfactuals and explanation. *Mind*, 115 (458):261–310, 2006.

Lewis, D. *Counterfactuals*. Blackwell, Oxford, 1973. ISBN 1-118-69641-7.

Lim, B. Y. and Dey, A. K. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, pp. 195–204. ACM, 2009.

Malle, B. F. Folk explanations of intentional action. *Foundations of social cognition*, 2011.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.

Pearl, J. *Causality*. Cambridge University Press, Cambridge, 2000.

Ribera, M. and Lapedriza, A. Can we do better explanations? a proposal of user-centered explainable ai. 2019.

Russell, C. Efficient search for diverse coherent explanations. *ACM FAT**, 2019.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. *ACM Conference on Fairness Accountability and Transparency*, 2019.

Various. 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`, 2018a.

Various. The What-If Tool. `https://pair-code.github.io/what-if-tool`, 2018b.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, forthcoming, 2018.

Weller, A. Challenges for transparency. *arXiv:1708.01870 [cs]*, 7 2017. URL `http://arxiv.org/abs/1708.01870`. arXiv: 1708.01870.