On the Strength of Causal Goodhart's Law

Adrien Majka [1] **Wassim Bouaziz** [1] **El-Mahdi El-Mhamdi** [1]

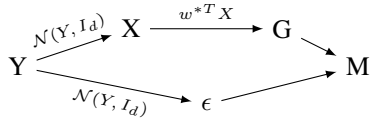
Abstract

Goodhart's law is an adage in policy-making stating that "*when a measure becomes a target, it ceases to be a good measure*". In the past few years, efforts have been made to formalise the law and assess its validity in the context of machine learning. Specifically, formalisms were proposed to distinguish cases where optimising a proxy metric is useful for an (unknown) intended goal from those where doing so harms the true goal. In the broader effort to formalise Goodhart's law, one central question is that of causality. Namely, does learning on a causal structure without being aware of it (and taking it into account) leads to a misalignment between the true goal and the proxy metric being optimised? This paper provides a positive answer to this question and proposes a causal formalism that separates three different causal relationships: (1) the classic case of a confounding factor, (2) a new causal structure we call the "*mirror confounding*", and (3) the cascading structure that we adapt from a previous work. Each causal structure involves a true goal, a proxy metric, the covariates on which the model learns and, when applicable, hidden variables.

1. Context and problem formulation

We follow the now common setup to formalise Goodhart's law (Manheim & Garrabrant, 2018; El-Mhamdi & Hoang, 2024): An unknown goal G is being optimised through a (known) proxy measure M that has, a priori, some resemblance with G, e.g., M is (assumed to be) well-correlated with G. The noise between the true goal G and the proxy metric M is captured by a quantity ϵ; depending on the causal structure, the influence of the noise ϵ can have different forms, and a hidden variable, denoted Y, can be a common cause to the covariates X, the goal G and the noise ϵ. Each of these quantities, summarised in Table 1, are linked by a causal graph. We consider three different settings which we formally specify in the following subsections.

Figure 1: Confounding factor schemes



1.1. Confounding factor

We consider the model illustrated in Figure 1, where a hidden variable Y determines at the same time the covariates used for prediction X and the noise ε. The true goal G is a direct product of the covariates. The proxy metric M is a sum of the true goal G and the noise term ε. In this setting, optimising for the noisy proxy metric M can lead to issues and misalignment with the true goal G.

Illustrative example. Confounding has already been considered in classical statistics (Splawa-Neyman et al., 1990; Street, 1990), in epidemiology (Greenland & Robins, 1986; Rubin, 1974) (see (Morabia, 2011) for a general history of the notion in epidemiology) and, recently, in the context of ML fairness (Plecko & Bareinboim, 2022; Schröder et al., 2023; Ashurst & Weller, 2023).

An example of confounding can be found in is credit score ratings (Dwork et al., 2011; Plecko & Bareinboim, 2022). The goal in this example is to determine the chances for an individual to default on their credit. The model has access

Quantity	Notation
Intended goal	G
Proxy metric	M
Covariate	X
Noise	ϵ
Hidden variable	Y
Goal characteristics	$\tilde{G}$

Table 1: Summary of the notation

[1]Center for applied Mathematics of Polytechnique (CMAP), Ecole Polytechnique, Palaiseau, France. Correspondence to: Adrien Majka <adrien.majka@polytechnique.edu>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment

Figure 2: Mirror confounding schemes

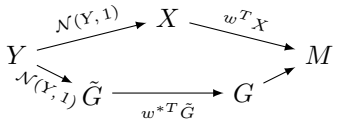
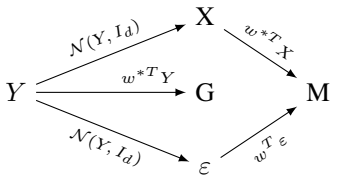


Figure 3: Cascading schemes



to several covariates (X in our model) that are supposed to be linked to credit default. The ground truth probability of default of every borrower would be the ideal goal to predict (G in our model, the true goal). This is, of course, never accessible. The training of the model must be held on a proxy metric (M in our model) only correlated to G. For credit score, historic decisions of credit attributions by bank agency and actual defaults for credit that are delivered is the proxy metric. However, the noise that differentiates the proxy metric from the true goal (ε in our model) can be correlated to the covariates by a hidden variable (Y). Identifying the true relationship of the covariates with the true goal can be difficult in such a setting. In the credit score example, the color of the credit applicant's skin - a protected attribute (Andreeva et al., 2004; Mehrabi et al., 2019) - is a hidden variable that influences the covariates (wealth, social status ...) as well as the metric used to train the algorithm (credit attribution).

1.2. Mirror confounding

We propose a scheme that we name "mirror confounding" due to its resemblance to the confounding scheme. The hidden variable Y determines the covariates X and a hidden vector $\tilde{G}$ that determines the goal G. The true goal is $G := w^{*T}\tilde{G}$. The metric is the sum of the true goal G and the covariates X multiplied by a term w, i.e : $M := G + w^T X$. Graph 2 summarises the model.

There is no causal link between the covariates X and the true goal G in this setting. The hidden variable Y is the only indirect link between the true goal G and the covariates.

Illustrative example. This situation would happen in the credit default setting if the covariates (X) on which the algorithm learns are linked to the risk of default (G) only through a third hidden variable (Y). The historic of credit reimbursement of a borrower would be such a covariate. It is positively correlated with the probability of reimbursement today G but only through a third hidden variable Y - the financial situation of the client.

1.3. Cascading

The hidden variables Y determines the covariates X, the true goal G and the noise ε. The true goal is $G := w^{*T}Y$. The metric M is the sum of the covariates X multiplied by w^{*T} and the noise ε multiplied by w, i.e : $M := w^{*T}X + w^T\varepsilon$. Graph 3 summarises the model.

We call this situation cascading as the hidden variable Y directly determines the goal G and is a direct antecedent of the covariates X and the noise term ε. However, the metric considered here does not directly contain the true goal G and is simply correlated to it via the effect of the hidden variables Y on the covariates X and the noise ε.

Illustrative example. Consider we want to predict the quality of a product from a noisy observation of its characteristics. A product has intrinsic (unknown) characteristics (the hidden variable Y) that directly determine its (unobserved) true quality (G). The intrinsic characteristics (Y) can be observed only in a noisy way (that would be X in our model). The product's price on the market is accessible and correlated to product's true quality. It is determined by the noisy observation of the characteristics (X) and some noise (ε). This noise (ε) is also influenced by the true characteristics of the product (Y). Using the price (M) as a proxy metric for product quality (G) would fit into the scheme we describe.

Contributions. The key question we address is the following. Does learning on a causal structure without taking it into account lead to a misalignment between the true goal and the proxy metric being optimised?

We provide a positive answer to this question as follows. We propose a causal formalism enabling an empirical study with 3 different causal relationships following an alignment setting : a true goal is approximated by a correlated proxy metric, with covariates on which the model learns and hidden variables. Our experiments show three consistent outcomes.

- Different causal structures lead to different outcomes in terms of discrepancy, showing that the underlying causal structure might be instrumental in identifying and alleviating a potential alignment problem.
- Increasing the sample size alleviates the discrepancy between the true goal and the proxy metric in the con-

founding case, but not in the case of mirror confounding and cascading.

- The three causal structures we studied are all leading to a form of Goodhart's law, stronger in the case of the cascading causal structure. They do not, however, completely prevent learning.

2. Experiment

2.1. Experimental setting

We chose to simulate the situation with Gaussian random variables to keep the experimental setting easy and interpretable. It allows us to keep a tight control of every quantity that might come into play in the learning setting. Moreover, we know the Gaussian setting to be an *easy* case for the alignment problem (El-Mhamdi & Hoang, 2024). In such a simple setting any complexity arising cannot be blamed on the complexity of the underlying probability distribution. This would not be the case for more complex situations such as power laws and other heavy-tailed distributions.

Confounding factor experiment. The hidden variable Y follows a standard Gaussian distribution with identity covariance matrix. The covariates X are drawn from a Gaussian distribution of mean equal to Y and identity covariance matrix, i.e $X \mid Y \sim \mathcal{N}(Y, I_d)$. The noise is drawn in a similar fashion from a Gaussian distribution of mean equal to Y and a covariance matrix equal to a constant times the identity, i.e $\varepsilon \mid Y \sim \mathcal{N}(Y, CI_d)$. This means marginally, we have $X \sim \mathcal{N}(0, 2I_d)$ and $\varepsilon \sim \mathcal{N}(0, (1+C)I_d)$. We have the true goal $G := X + w^T\varepsilon$. w^{*T} and w are drawn from a uniform distribution over $[0,1]^d$ and then fixed. They do not vary across samples.

Mirror confounding experiment. The vector of characteristics $\tilde{G}$ follows a Gaussian distribution centered in Y with identity covariance matrix, which means that unconditionally on Y we have $\tilde{G} \sim \mathcal{N}(0, 2I_d)$. Then we have $G := w^{*T}\tilde{G}$, where $w^* \sim U[0,1]^d$. The covariates X are drawn from a Gaussian distribution centered in Y with identity matrix times a constant C, which implies $X \sim \mathcal{N}(0, 2I_d)$. The metric is $M := w^T X + G$, with $w \sim U[0,1]^d$. Both $w^* \sim U[0,1]^d$ and $w \sim U[0,1]^d$ are drawn at the beginning of the experiment and do not vary from each sample.

Cascading experiment. The hidden variable Y is drawn following a centered Gaussian random variable with identity covariance matrix. The true goal G is deduced from it with the relation $G := w^{*T}Y$, where $w^* \sim U[0,1]^d$ is drawn at the beginning of the experiment and does not vary from each sample. The vector of covariates X follows a Gaussian random variable centered in Y, which means that unconditionally on Y we have $X \sim \mathcal{N}(0, 2I_d)$. The noise ε is drawn from a Gaussian random variable centered in Y with identity matrix times a constant C, which implies $\varepsilon \sim \mathcal{N}(0, (1+C)I_d)$. The metric M is linked to the covariates by the same coefficient w^* as the hidden variable X and the true goal M. The noise vector is multiplied by $w \sim U[0,1]^d$ (which is drawn at the beginning of the experiment and does not vary from each sample). This gives $M := w^{*T}X + w^T\varepsilon$

2.2. Experimental results

We trained 1440 models, spanning 4 values of noise to signal ratio in the metric (1, 0.1, 0.01, 0.001 and 0.0001), 3 sample size values with 4 data seeds and 10 model seeds. The covariates (X) are in dimension 10 and the learning rate is constant at 0.001. The model we train is a multilayer perceptron (MLP) with an initial linear layer from dimension 10 to 64, followed by 3 hidden layers of width 64 with ReLU activation and a final linear layer to combine all of it.

Figure 4 shows aggregated trajectories with the same noise variance in the three aforementioned cases. In the confounding case, after a brief decrease of the true loss value at the beginning of the optimisation, the true loss value rises and stays at a higher value than early in the optimisation process.

Mirror confounding and cascading follow a different path as the true loss decreases first to then stabilise. In the confounding case, the true loss value is an order of magnitude higher than the training loss. In cascading and mirror confounding, it is roughly a factor of twenty.

Figure 6 shows the discrepancy between the true loss and the proxy metric as the sample size increases from 100 to 10000. In the case of mirror confounding, the discrepancy decreases from 0.25 to 0.20 with increasing sample size. For cascading, after the discrepancy decreased from 0.325 to 0.225 when augmenting the sample size from 100 to 1000, it decrease at 0.20 as we increase the sample size to 10000. For confounding, the discrepancy is stable for the 2 first sample size values around 0.20 and the decrease to 0.125.

In the three cases, the discrepancy between the proxy metric and the true goal augment with the noise to signal ratio. In the case of confounding and mirror confounding, the discrepancy is slightly negative for value of the noise to signal ratio inferior to 10^{-2}. For cascading, the discrepancy decrease when diminishing the noise to signal ratio from 1 to 10^{-1}, and then stabilises at different level for each sample size.

Figure 4: Train losses and true loss (goal) for different schemes

(a) Confounding (b) Mirror confounding (c) Cascading

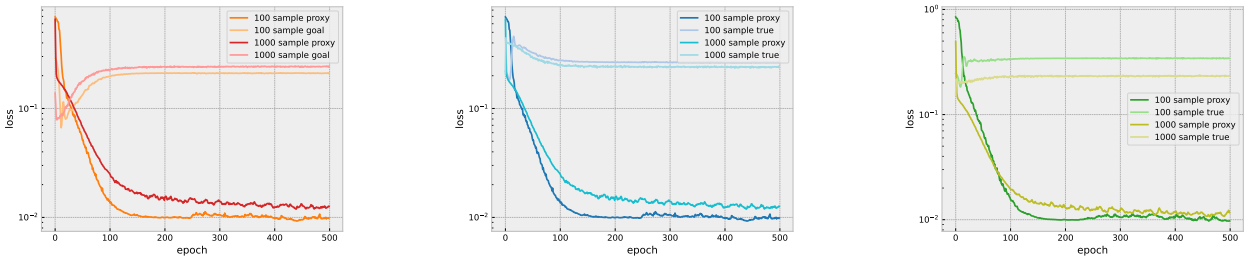


Figure 5: Train losses and true loss (goal) discrepancy for different schemes and varying noise to signal ratio.

(a) Confounding (b) Mirror confounding (c) Cascading

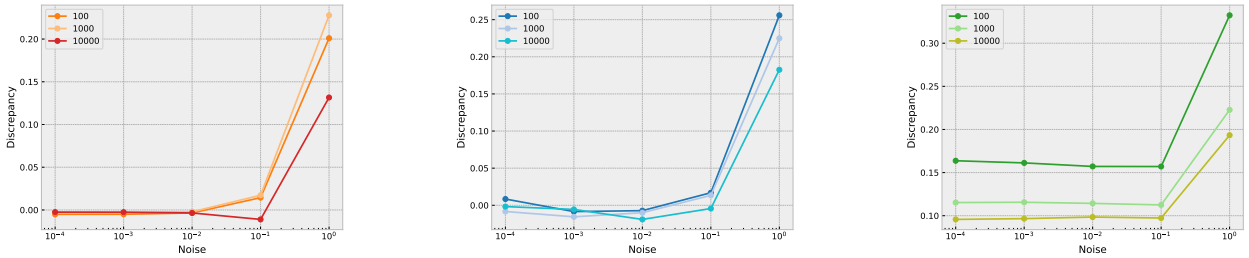
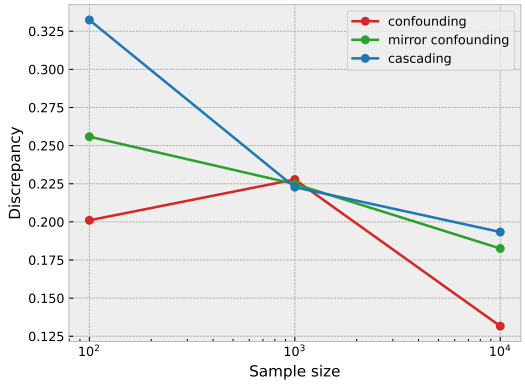


Figure 6: Values of the discrepancy between the proxy metric and true goal according to the sample size for a constant noise to signal ratio (equal to 1)



3. Discussion

In this work, we ask the question, "*does learning on a causal structure without taking it into account leads to a misalignment between the true goal and the proxy metric being optimised?*" which we answer positively. Further analysis should address the following two questions.

Causal Structure importance. Our experiments show that different causal structures imply different final discrepancies between the true goal and the proxy metric. For example, cascading has a true goal roughly twenty times higher than its proxy metric loss in Figure 4. Looking at Figure 5, in cascading, even with decreasing noise to signal ratio, the discrepancy between the true goal and the proxy metric is constant. Getting theoretical results linking the property of the causal graph to the true goal and proxy metric discrepancy would be a major landmark in understanding such results.

Learning dynamic. We can observe that the algorithm learns at least a little bit of the goal even in the most "hostile" settings where the metric is not a direct product of the goal (see Figure 4). In the cascading and mirror confounding cases, the true goal loss decreases similarly to the proxy

metric at the beginning of the training procedure. In the confounding case, the algorithm seems to actively *unlearn* the true goal after a few epochs. Experimenting in other causal settings could allow us to have a more global image of the influence of the causal structure on the alignment problem. Also, transposing the setup to a real-world dataset is of prime importance but also poses important challenges, as most of the time the true goal is not accessible and hence must be very carefully crafted.

References

Andreeva, G., Ansell, J., and Crook, J. Impact of anti-discrimination laws on credit scoring. *Journal of Financial Services Marketing*, 9(1):22–33, September 2004. ISSN 1363-0539, 1479-1846. doi: 10.1057/palgrave.fsm.4770138. URL https://link.springer.com/10.1057/palgrave.fsm.4770138.

Ashurst, C. and Weller, A. Fairness Without Demographic Data: A Survey of Approaches. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–12, Boston MA USA, October 2023. ACM. ISBN 979-8-4007-0381-2. doi: 10.1145/3617694.3623234. URL https://dl.acm.org/doi/10.1145/3617694.3623234.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness, 2011. URL https://arxiv.org/abs/1104.3913. Version Number: 2.

El-Mhamdi, E.-M. and Hoang, L.-N. On Goodhart's law, with an application to value alignment, 2024. URL https://arxiv.org/abs/2410.09638. Version Number: 1.

Greenland, S. and Robins, J. M. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, 15(3):413–419, 1986. ISSN 0300-5771, 1464-3685. doi: 10.1093/ije/15.3.413. URL https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/15.3.413.

Manheim, D. and Garrabrant, S. Categorizing Variants of Goodhart's Law, 2018. URL https://arxiv.org/abs/1803.04585. Version Number: 4.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A Survey on Bias and Fairness in Machine Learning, 2019. URL https://arxiv.org/abs/1908.09635. Version Number: 3.

Morabia, A. History of the modern epidemiological concept of confounding. *Journal of Epidemiology & Community Health*, 65(4):297–300, April 2011. ISSN 0143-005X. doi: 10.1136/jech.2010.112565. URL https://jech.bmj.com/lookup/doi/10.1136/jech.2010.112565.

Plecko, D. and Bareinboim, E. Causal Fairness Analysis, 2022. URL https://arxiv.org/abs/2207.11385. Version Number: 1.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, October 1974. ISSN 1939-2176, 0022-0663. doi: 10.1037/h0037350. URL https://doi.apa.org/doi/10.1037/h0037350.

Schröder, M., Frauen, D., and Feuerriegel, S. Causal Fairness under Unobserved Confounding: A Neural Sensitivity Framework. 2023. doi: 10.48550/ARXIV.2311.18460. URL https://arxiv.org/abs/2311.18460. Publisher: arXiv Version Number: 3.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1990. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/2245382. Publisher: Institute of Mathematical Statistics.

Street, D. J. Fisher's Contributions to Agricultural Statistics. *Biometrics*, 46(4):937, December 1990. ISSN 0006341X. doi: 10.2307/2532439. URL https://www.jstor.org/stable/2532439?origin=crossref.