Strategyproof Explainable AI

Thomas W. L. Norman*

Magdalen College, Oxford

January 5, 2022

Abstract

Artificial intelligence is an increasingly common tool for market competition, producing algorithms that are often difficult to analyze and explain. The field of 'explainable AI' offers a promising response, but the popular 'additive feature attribution' methods (LIME and SHAP, for instance) may be vulnerable to manipulation hiding an algorithm's true nature. Nevertheless, I offer here tax schemes under which it is suboptimal to engage in such manipulation. As well as being strategyproof in this way, taxing via Shapley values is distinguished from other schemes by virtue of an attractive efficiency property. *Journal of Economic Literature* Classification: C71, D62, L40

Key Words: black box explanations, model interpretability, cost sharing, strategyproof mechanisms, Shapley values

1 Introduction

There is an increasing tendency for firms to employ artificial intelligence (AI) as a competitive tool, with the associated algorithms also becoming increasingly complex and inscrutable. The nascent field of 'explainable AI' quantifies the importance to an algorithm's output of its various constituent inputs (or 'features'). It thus seeks to open the 'black box' of machine learning algorithms, and represents a promising toolkit for regulation of their use in markets. However, there is concern that such tools might be manipulable by the algorithm designer to hide particular features, as Slack et al. (2020) have demonstrated for popular 'additive feature attribution' techniques such as LIME and SHAP, the latter of which weights explanatory features according to their Shapley values.

Nevertheless, even if such manipulation is possible, it need not be in a firm's interests under an appropriately designed tax scheme. Here I analyze a general market setting, where the use of algorithms generates external costs that the social planner seeks to recover from a set of firms in Pigouvian fashion. A topical example is offered by the

*Email thomas.norman@magd.ox.ac.uk.

concern over pricing algorithms that may learn to behave anticompetitively (see, e.g., Calvano et al. 2020). This fits within the framework of the classic cost sharing problem in cooperative game theory (Young 1994), and in particular is a more-or-less direct application of the Moulin and Shenker (2001) model.

I use this model to show that sharing external costs among an algorithm's constituent features according to their Shapley values is 'strategyproof', in the sense of inducing truthful behavior from (any coalition of) firms. Whilst a broad class of 'feature-additive' models also resists manipulation in this way, the approach of sharing costs according to Shapley values is uniquely attractive from an efficiency perspective, in the sense of risking the lowest maximal welfare loss.

2 The Model

Suppose that a market consists of a set $\{1, \ldots, J\}$ of firms employing *algorithms* $f \equiv (f^1, \ldots, f^J)$, with each f^j mapping from some metric space X^j of *inputs* into a field F^j of *outputs*; let $X \equiv \prod_{j=1}^{J} X^j$ and $F \equiv \prod_{j=1}^{J} F^j$. For instance, algorithms for oligopolistic competition may produce a list of prices based on various market data, or a list of quantities, or a list of private and external costs and benefits of interested parties. Since I will be concerned only with the problem of a welfare-maximizing regulator, I can take the relevant output field to be the set of possible *external costs* arising from the algorithms' use, $F^j = \mathbb{R}_+$, $j = 1, \ldots, J$—for instance, the costs to consumers arising from any anticompetitive features of pricing algorithms.

Now, the mapping $f : X \to F$ may be sufficiently complex that there is value in a simpler *explanation model* $c : \{0,1\}^N \to F$ of the whole market, which uses a set of N binary variables called *(simplified input) features* $z \in \{0,1\}^N$ related to the original inputs by some mapping function $x = h_x(z)$. For Shapley values, for instance, h_x maps 1 or 0 to the original input space to indicate that the input is included in the model or not. A model c is a *local explanation* if it is designed to explain an output $f(x)$ based on local perturbations of a single input x, as with LIME (Local Interpretable Model-agnostic Explanations, Ribeiro, Singh, and Guestrin 2016) and SHAP (SHapley Additive exPlanation, Lundberg and Lee 2017). These are also examples of what Lundberg and Lee call *additive feature attribution methods*; they produce explanation models that are linear functions of simplified input features:

$$c(z) = \phi_0 + \sum_{i=1}^{N} \phi_i z_i,$$

where $\phi_i \in \mathbb{R}$ is the *effect* attributed to feature i (the sum of all of which approximates

the original algorithms' output $f(x)$) .[1]

Given $z \in \{0,1\}^N$, let $Z \equiv \{i \,|\, z_i = 1\}$ record the set of (indices of) features *included in the explanation model* (i.e. those with binary value 1). To find $\phi \equiv (\phi_0, \ldots, \phi_N)$, LIME minimizes the sum of the model c's mean squared error and its complexity; SHAP, meanwhile, estimates the features' Shapley (1953) values,[2]

$$\xi_i^* = \sum_{S \subseteq Z \setminus i} \frac{|S|!(|Z| - |S| - 1)!}{|Z|!} \left[C(S \cup i) - C(S)\right], \quad i \in Z,$$

where $C(S) \equiv c(z_S)$, $S \subseteq Z$, gives the external cost from applying the model c to the binary vector $z_S \in \{0,1\}^N$ which has 1's for the set of features S and 0's otherwise.

Suppose for simplicity that, for each feature $i = 1, \ldots, N$, there is at most one firm $j(i)$ with a nonzero *willingness to pay* $u_i \in \mathbb{R}_+$ to *include feature* i, i.e. to employ an algorithm with an input belonging to the h_x-image of $\{z \in \{0,1\}^N \,|\, z_i = 1\}$; thus $J \leq N$, with each firm allowed to value multiple features. The regulator may decide the subset Z of features included in the explanation model, and charge a tax to any firm $j \in J$ whose algorithm includes a feature in Z. Its motivation in so doing is to maximize social welfare, composed for each feature $i = 1, \ldots, N$ of firm $j(i)$'s willingness to pay u_i to include feature i (net of the associated tax) and the resulting (approximate) external cost. I henceforth assume C to be nondecreasing, $S \subseteq T \Rightarrow C(S) \leq C(T)$, which implies that each ϕ_i is nonnegative. This is obviously a restrictive assumption in the context of pricing algorithms, for instance, where some features may be procompetitive and hence reduce external costs; however, the focus of the analysis here is on externally harmful features, and the scope of the exercise below can be confined to these.

The function C is *submodular* if the marginal cost $C(S \cup i) - C(S)$ of adding feature i to a certain set S of other features does not increase when the set S expands. Thus, if C were a symmetrical function of each feature, a submodular C would be a concave function of $|S|$. Since, for any $i \in Z$ and any $S \subseteq Z$, $C(S \cup i) - C(S) = \phi_i z_i$, submodularity of C clearly holds for an additive feature attribution method. I assume that, whilst C is known to the regulator, the u_i's are not. A *cost sharing method* is a function ξ assigning a nonnegative cost share $\xi_i(S)$ to each feature $i \in S$, in a manner that satisfies *budget balance*, $\xi_S(S) = C(S)$; ξ is *cross-monotonic* if feature i's cost share cannot increase when the set of included features expands:

$$S \subseteq T,\ i \in S \quad \Rightarrow \quad \xi_i(T) \leq \xi_i(S).$$

It is natural to consider a mechanism that elicits from each firm its willingness

1. For pricing algorithms, for instance, relevant features might include whether there has been a recent price cut in the market, or a change in the cost of production.

2. See Lundberg and Lee (2017) for details.

to pay for each feature $i = 1, \ldots, N$, then decides which features Z are included in the explanation model and how the external cost $C(Z)$ is to be shared among those features. More specifically, a *revelation mechanism* is a mapping M assigning to each profile $u \in \mathbb{R}^N_+$ a subset $Z(u) \subseteq \{1, \ldots, N\}$ of features included in the explanation model (or equivalently, the vector $z(u)$ with $z_i(u) = 1$ for all $i \in Z(u)$ and $z_i(u) = 0$ for all $i \notin Z(u)$) and a vector $\tau(u) \in \mathbb{R}^N_+$ of taxes. For any feature $i = 1, \ldots, N$, firm $j(i)$ is assumed to have additively separable quasi-linear utility in i, $u_i z_i - \tau_i$. For the purposes of Proposition 1 below, this quasilinearity assumption is unrestrictive, capturing any preferences that are strictly increasing in money, but for Proposition 2 it is essential (see Moulin and Shenker 2001, Comment 1).

Given any subset $S \subseteq \{1, \ldots, N\}$ of features and any two profiles u, u' such that $u_i = u'_i$ for all $i \notin S$, let (z, τ) and (z', τ') denote the allocations implemented by M at u and u' respectively. Then *group strategyproofness* of M requires that

$$\{\forall i \in S : u_i z'_i - \tau'_i \geq u_i z_i - \tau_i\} \quad \Rightarrow \quad \{\forall i \in S : u_i z'_i - \tau'_i = u_i z_i - \tau_i\}.$$

In words, if no feature in S yields lower benefit to the relevant firm by changing from u to u', then nor does any feature in S yield higher benefit. This is a strong form of strategyproofness that implies in particular that: (a) no firm j has an incentive to misrepresent its benefit from the inclusion of any features in the explanation model (what might be called *firm strategyproofness*); and (b) no group of firms has a joint incentive to misrepresent their benefits from the inclusion of any subset of features.

There are a number of other properties that I will require of a mechanism; they are those of Moulin and Shenker (2001), adapted to the current setting:

No Positive Transfers (NPT): Each $i = 1, \ldots, N$ has a nonnegative cost share, $\tau_i \geq 0$.

Voluntary Participation (VP): The benefit derived from noninclusion ($z_i = 0$) of feature i at no cost ($\tau_i = 0$) is guaranteed to firm $j(i)$ if it reports truthfully.

Firm Sovereignty (FS): For each feature $i = 1, \ldots, N$, firm $j(i)$ has a message u_i guaranteeing that the feature is included in the explanation model ($z_i = 1$), regardless of the values u_{-i} reported for other features.

These are standard assumptions in the cost sharing literature, and are discussed further by Moulin and Shenker.

Given a cost sharing method ξ and the willingness to pay profile $u \equiv (u_i)_{i=1}^N$, suppose that for each feature $i = 1, \ldots, N$, firm $j(i)$ decides whether or not to request that feature i be included in the explanation model ($z_i = 0$ or $z_i = 1$)—or, equivalently, that the h_x-image of $\{z \in \{0,1\}^N \mid z_i = 1\}$ be included in the algorithms—and taxes τ are then levied on the included features according to the cost sharing method ξ. Is it optimal for firms to behave truthfully in this *demand game*? Misrepresentation of u_i

could be accomplished through Slack et al.'s (2020) creation of 'scaffolding' around the algorithm's output to downplay the explanatory role of a particular feature. Moulin and Shenker (2001) show that, if ξ is cross-monotonic, then the demand game has a unique strong equilibrium $\tilde{Z}(\xi, u)$ (Aumann 1959) that induces a revelation mechanism $M(\xi)$ that is group strategyproof:[3]

$$Z(u) = \tilde{Z}(\xi, u);\ \tau_i(\tilde{Z}(\xi, u)) \text{ if } i \in \tilde{Z}(\xi, u);\ \tau_i(u) = 0 \text{ otherwise.} \tag{1}$$

Lemma 1 (Moulin and Shenker 2001) *For any submodular C and cross-monotonic cost sharing method ξ, the mechanism $M(\xi)$ (defined by (1)) is budget balanced, meets NPT, VP, FS, and is group strategyproof.*

Since the Shapley value is cross-monotonic when C is submodular (see, e.g., Sprumont 1990), the following is immediate.

Corollary 1 (Moulin and Shenker 2001) *The cost sharing method ξ^* derived from the Shapley value additive feature attribution method is group strategyproof.*

In fact, we can construct a group strategyproof cost sharing method from any additive feature attribution method, including LIME: Say that the cost sharing method is *feature-additive* if, for any $S \subseteq Z$ and any feature $i \in S$,

$$\xi_i(S) = \frac{\phi_i}{\sum_{k \in S} \phi_k} C(S).$$

Thus, a feature-additive cost sharing method shares costs in proportion to features' effects in the explanation model.

Proposition 1 *Given an additive feature attribution method, if ξ is a feature-additive cost sharing method, then the mechanism $M(\xi)$ (defined by (1)) is budget balanced, meets NPT, VP, FS, and is group strategyproof.*

Proof. Suppose otherwise; then, by Lemma 1, ξ is not cross-monotonic:

$$\begin{aligned} \exists S \subseteq T,\ i \in S \quad &\Rightarrow \quad \frac{\phi_i}{\sum_{k \in T} \phi_k} C(S) > \frac{\phi_i}{\sum_{k \in S} \phi_k} C(T) \\ &\Leftrightarrow \quad \phi_i \left(C(T) \sum_{k \in T} \phi_k - C(S) \sum_{k \in S} \phi_k \right) < 0, \end{aligned}$$

contradicting the assumption that C is nondecreasing. ∎

3. Note that the classic individual strategyproofness result of Dasgupta, Hammond, and Maskin (1979) is insufficient even for firm strategyproofness here, since it would apply to individual features rather than each firm's collection of features.

Whilst there are many cross-monotonic cost sharing methods that satisfy the conditions of Proposition 1, Moulin and Shenker (2001) show that the Shapley value has an additional attractive feature, not shared by feature-additive cost sharing methods.[4] Given a cost sharing method ξ, let $\gamma(\xi, u)$ be the welfare loss under profile u and let $\gamma(\xi)$ be the maximal welfare loss of the mechanism $M(\xi)$ defined by (1):

$$\gamma(\xi) \equiv \sup_{u \in \mathbb{R}_+^N} \gamma(\xi, u) = \sup_{u \in \mathbb{R}_+^N} \left[w(\{1, \ldots, N\}, u) - (u_{\tilde{Z}(\xi,u)} - C(\tilde{Z}(\xi, u))) \right],$$

where $w(S, u) \equiv \max_{T \subseteq S}[u_T - C(T)]$ and $u_T \equiv \sum_{i \in T} u_i$.

Proposition 2 (Moulin and Shenker 2001) *Among all mechanisms $M(\xi)$ derived from cross-monotonic cost sharing methods, that which allocates costs according to Shapley values has the uniquely smallest maximal efficiency loss $\gamma(\xi)$.*

4. This is over and above the properties of local accuracy, missingness and consistency that have already been established for the Shapley value approach (see Lundberg and Lee 2017, Theorem 1, based on Young 1985), as well as Shapley's (1953) original properties of course.

References

Aumann, R. J. 1959. "Acceptable Points in General Cooperative n-Person Games." In *Contributions to the Theory of Games IV,* Annals of Mathematics Study 40*,* edited by R. D. Luce and A. W. Tucker, 287–324. Princeton, NJ: Princeton University Press.

Calvano, E., G. Calzolari, V. Denicolò, and S. Pastorello. 2020. "Artificial Intelligence, Algorithmic Pricing, and Collusion." *American Economic Review* 110:3267–3297.

Dasgupta, P., P. Hammond, and E. Maskin. 1979. "The Implementation of Social Choice Rules." *Review of Economic Studies* 46:185–216.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Neural Information Processing Systems (NIPS),* edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774. Curran Associates, Inc.

Moulin, H., and S. Shenker. 2001. "Strategyproof Sharing of Submodular Costs: Budget Balance versus Efficiency." *Economic Theory* 18:511–533.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." Working paper, arXiv 1602.04938v3.

Shapley, L. S. 1953. "A Value for n-Person Games." In *Contributions to the Theory of Games II,* edited by H. W. Kuhn and A. W. Tucker, 307–317. Princeton: Princeton University Press.

Slack, D., S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. 2020. "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods." Working paper, arXiv 1911.02508v2.

Sprumont, Y. 1990. "Population Monotonic Allocation Schemes for Cooperative Games with Transferable Utility." *Games and Economic Behavior* 2:378–394.

Young, H. P. 1985. "Monotonic Solutions of Cooperative Games." *International Journal of Game Theory* 14:65–72.

———. 1994. "Cost Allocation." In *Handbook of Game Theory with Economic Applications,* edited by R. J. Aumann and S. Hart, 2:1193–1235. Amsterdam: Elsevier.