Explanation Design in Strategic Learning: Sufficient Explanations That Induce Non-harmful Responses

Kiet Q. H. Vo[1] **Siu Lun Chau**[2] **Masahiro Kato**[3]
Yixin Wang[4] **Krikamol Muandet**[1]
[1]Rational Intelligence Lab, CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[2]College of Computing & Data Science, Nanyang Technological University, Singapore
[3]Mizuho–DL Financial Technology, Co., Ltd., Tokyo, Japan
[4]University of Michigan, Ann Arbor, MI, USA

Abstract

We study the design of explanations in algorithmic decision-making with strategic agents—individuals who may modify their inputs in response to explanations of a decision maker's (DM's) predictive model. While the demand for algorithmic transparency has led much prior work to assume full model disclosure, in practice DMs typically provide only partial information via explanations, which can cause agents to misinterpret the model and take actions that unintentionally reduce their own utility. A central open question is therefore how DMs should communicate explanations that avoid harming strategic agents while still supporting their own goals, e.g., minimising predictive error. In this work, we analyse widely used explanation methods and establish a necessary condition to prevent explanations from inducing self-harming responses. Furthermore, we show that *action recommendation-based explanations* (ARexes), which encompass counterfactual explanations, are sufficient to induce all non-harmful responses. Under a conditional homogeneity assumption, this sufficiency extends to ARex-generating methods, echoing the revelation principle in information design. To demonstrate their practical utility, we introduce a simple learning procedure that jointly optimises the predictive model and the explanation-generating policy. Experiments show that ARexes enable DMs to achieve high predictive performance while preserving agents' utility, offering a principled strategy for safe and effective partial model disclosure.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Modern regulatory frameworks emphasise transparency in algorithmic decision making, mandating that decision makers (DMs) provide clear justifications for automated decisions (Selbst and Powles, 2017; Wachter et al., 2017a). For example, the General Data Protection Regulation (GDPR) (Council of European Union, 2016) includes provisions commonly referred to as the *right to explanation*, requiring DMs to inform agents (i.e., individuals affected by DMs' decisions) about the basis of these decisions in a comprehensible manner (Goodman and Flaxman, 2017). These provisions aim to help agents both understand and contest algorithmic decisions. However, transparency can incentivise agents to manipulate their inputs to secure more favorable outcomes, triggering strategic adaptations by both agents and DMs (Hardt et al., 2016). This dynamic has motivated extensive research into modeling strategic behavior and optimising decision making under such interactions (Miller et al., 2020). Within this line of work, explainability is often equated with full disclosure of the decision making model, including its architecture and parameters (Shavit et al., 2020; Harris et al., 2022b; Vo et al., 2024). Such disclosure ensures transparency by enabling agents to simulate and assess alternative actions.

Although full disclosure may formally satisfy transparency requirements, it often pushes DMs towards adopting inherently interpretable models at the expense of predictive performance (e.g. linear models). For complex models with billions of parameters, full disclosure rarely translates into genuine interpretabil-

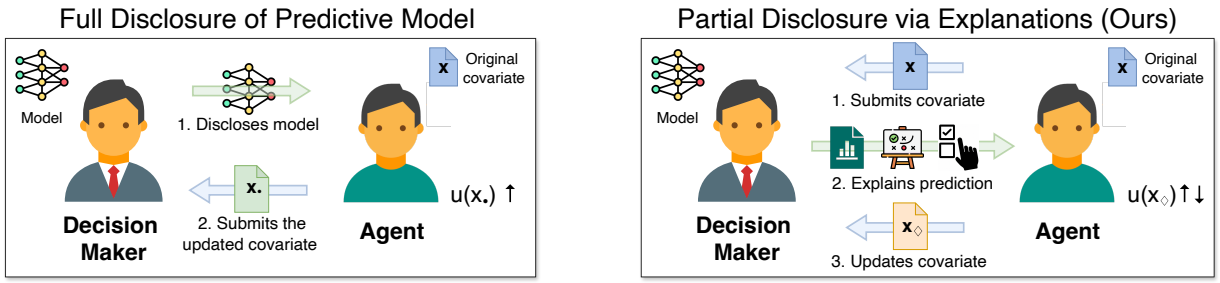


Figure 1: (left) With full access to the DM's model, the agent with a covariate x can *correctly* anticipate how changes affect predictions and choose a response $x_\bullet$ that reliably improves utility $u(x_\bullet)$. (right) With only an explanation, the agent's response $x_\diamond$, based on partial information, might not improve utility $u(x_\diamond)$.

ity for agents and may instead overwhelm them with excessive detail. This raises important questions about whether full disclosure truly serves the original intent of the right to explanation. Moreover, it may conflict with the interests of DMs, especially when sensitive intellectual property is involved. For example, in car insurance pricing, mandating full disclosure could expose a company's proprietary model to competitors, who could exploit it without bearing the development costs, eroding the firm's competitive edge.

In practice, explaining a DM's predictive model does not necessitate full disclosure. Instead, conveying partial information can also be considered (Christoph, 2020). While many explanation methods have been studied in strategic learning (Tsirtsis and Gomez Rodriguez, 2020; Xie and Zhang, 2024; Cohen et al., 2024), it remains unclear which approach would best serve DMs. Specifically, when explanations omit certain details of the underlying model, agents may misinterpret the model's behaviour and take actions that result in suboptimal or even detrimental outcomes for them (e.g., see Appendix A.3). Popular explanation methods typically highlight certain model characteristics[1] rather than guiding agents' strategic responses (Lundberg, 2017; Tsirtsis and Gomez Rodriguez, 2020; Chau et al., 2022, 2023). As a result, explanations may inadvertently mislead agents into harmful actions, eroding trust and creating perceptions of unfairness or unreliability. This motivates a new challenge: designing explanations for strategic agents that prevent harm while supporting the DM's objectives across diverse decision-making settings. By analogy to information design (c.f. Section 5), we call this problem *Explanation Design*. As a result, our work addresses the following question:

When full disclosure of the predictive model is neither feasible nor desirable, what kind of explanations should be communicated to ensure transparency while balancing the interests of all parties involved?

Our contributions. We develop a theoretical framework for analysing explanation design in strategic learning, with four main contributions: (i) we show that, unlike agents with full access to the predictive model, those who rely solely on explanations may misinterpret the model's behaviour and take actions that reduce their utility. (ii) Under common assumptions on agents' behavior, we derive a necessary condition (Theorem 3.2 and Corollary B.1) for explanations to avoid misleading agents into utility-harming actions. In contrast, we show that counterfactual explanations (Wachter et al., 2017b), by design, do not mislead agents (Remark 3.3). (iii) We formalise the class of *action recommendation-based explanations* (ARexes, singular: ARex), which encompass counterfactual explanations and show that this class is sufficient to induce all non-harmful actions (Proposition 3.4). Under conditional homogeneity of agents' responses, this sufficiency extends to ARex-generating methods (Theorem 3.6), thus offering a principled framework for designing safe explanations in strategic learning. (iv) We demonstrate the practical value of ARexes in synthetic and real-world optimisation tasks, where the DM jointly learns the predictive model and ARex policy. Results show that ARexes enable the DM to improve predictive performance while preserving agents' utility. All proofs are provided in Appendix C.

2 PROBLEM FORMULATION

Notations. Random variables are denoted with uppercase letters (e.g., X) and their realisations with lowercase letters (e.g., x). The index set $\{1, \dots, T\}$ is denoted as $[T]$. We use $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ to denote the spaces of agents' observed covariates, outcomes, and unobservable, respectively. The model class and explanation space are respectively denoted by $\mathcal{G}$ and $\mathcal{E}$. We use double dot to denote the initial value of agent t's variable (e.g., $\ddot{x}_t$), as opposed to the value after it

[1] We provide a more detailed discussion contrasting the *epistemic* and *strategic roles* of explanations in Appendix B.1.

has been shifted, due to strategic behaviour (e.g., x_t).

We use the car insurance pricing (Shavit et al., 2020) as our running example and consider the scenario in which a DM (an insurer), interacts with a population of agents (customers), indexed by the integer t. Agents interact with the DM separately and independently. For simplicity, we describe the setup for a single agent. Let $(\ddot{X}, Z)$ be random variables jointly distributed as $P_{\ddot{X},Z}$. For each agent t, let $(\ddot{x}_t, z_t)$ denote an independent realisation of $(\ddot{X}, Z)$. Here, $\ddot{x}_t \in \mathcal{X}$ represents agent's observable features (e.g., driving records, car's model and features), while $z_t \in \mathcal{Z}$ captures unobservable factors (e.g., socioeconomic factors). Furthermore, the agent has an unrealised outcome $\ddot{y}_t := h(\ddot{x}_t, z_t)$, where $h : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is a deterministic potential outcome function and $\ddot{y}_t \in \mathcal{Y} \subseteq \mathbb{R}$. In our car insurance context, $\ddot{y}_t$ reflects the future accident cost of this customer, which the DM tries to predict to determine the insurance premium.

In the beginning, the DM selects a predictive model g from the hypothesis class $\mathcal{G} \subseteq \{g' : \mathcal{X} \to \mathcal{Y}\}$, to approximate h, e.g., by learning g from historical data. The agent first submits their base covariate $\ddot{x}_t$, receiving a preliminary prediction $\hat{\ddot{y}}_t := g(\ddot{x}_t)$. In addition, the DM provides an explanation $e_t \in \mathcal{E}$ describing how $\hat{\ddot{y}}_t$ was computed. In practice, this means a customer submits an insurance application, receives a quoted premium, and is shown an explanation of how that premium was determined. In this work, we formalise explanations and associated concepts as follows.

Definition 2.1. An *explanation method* is a tuple $(\mathcal{E}, \sigma)$ where $\mathcal{E}$ is the space of feasible explanations and $\sigma : \mathcal{X} \times \mathcal{G} \to \mathcal{E}$ is an *explanation policy* that picks an *explanation* $e \in \mathcal{E}$ for the agent with base covariate $x \in \mathcal{X}$ w.r.t. the predictive model $g \in \mathcal{G}$. The explanation e is a global explanation if σ is a constant function w.r.t. the input x. Otherwise, σ is said to generate local explanations.

This definition[2] allows us to incorporate a wide range of local and global explanation methods and analyse their impact on guiding agent behavior, as we explore in Section 3. The explanation space $\mathcal{E}$ is general and depends on the chosen explanation method. For example, when using a global surrogate model, such as a linear function, to approximate the predictive model g (Molnar, 2020), the explanation space $\mathcal{E}$ is a subset of linear functions $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y} : f(x) = w^\top x + b\}$. Alternatively, attribution-based methods such as SHAP (Lundberg, 2017) can have $\mathcal{E} \subseteq \mathbb{R}^d$ where each explanation e is a d-dimensional vector of feature attributions. Several other explanation methods fit within this framework, and we provide a more comprehensive discussion of them in Appendix A.1.

[2]We note that this definition is intentionally broad and places no restriction on the epistemic value of explanations (i.e., their ability to support model understanding). See Appendix A.1 for further discussion.

Agent's reaction. After receiving an explanation e_t, the agent seeks to improve their predictive score by strategically modifying their features before resubmitting (Tsirtsis and Gomez Rodriguez, 2020; Xie and Zhang, 2024; Karimi et al., 2021; Harris et al., 2022a). For example, an insurance customer might take actions such as installing a telematics device or upgrading to a safer vehicle so as to lower their predicted risk and premium when reapplying for a contract. These actions often come with tangible costs such as time, money, or effort. The agent must weigh the benefit of a more favourable prediction against the cost of implementing changes. Following standard practice in strategic learning (Hardt et al., 2016; Shavit et al., 2020; Harris et al., 2022b; Vo et al., 2024), we model this trade-off with an additive utility function $u_t : \mathcal{G} \times \mathcal{X} \to \mathbb{R}$:

$$u_t(g, x) := b(g, x) - c_t(\ddot{x}_t, x) := (-g(x)) - c_t(\ddot{x}_t, x), \quad (1)$$

where $b(g, x)$ denotes the potential benefit associated with the prediction $g(x)$, while c_t captures the cost of changing features from $\ddot{x}_t$ to x. In our example, the agent prefers lower prediction scores, hence $b(g, x) := -g(x)$. However, the framework straightforwardly generalises to settings where higher scores are desirable, e.g., in credit scoring.

We model heterogeneity in agents' cost functions by introducing a random function, i.e., function-valued random variable, $C \sim P_C$, drawn from a distribution P_C over a family of cost functions. We allow C to be statistically correlated with $\ddot{X}$ and Z. For example, a customer's ability to modify their features, e.g., upgrading the vehicle, may depend on personal characteristics such as income or socioeconomic background, captured in $\ddot{X}$ and Z. Each agent has a different cost function c_t: a realisation of C and unknown to the DM. We further define the cost function c_t as follows:

Definition 2.2 (Cost function)**.** A function $c_t : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a cost function for agent t if it satisfies $c_t(\ddot{x}_t, \ddot{x}_t) = 0$ and $c_t(\ddot{x}_t, x) > 0$ for all $x \neq \ddot{x}_t$.

In classical strategic learning (Hardt et al., 2016), an agent with full knowledge of g chooses a *best response* x that maximises $u_t(g, x)$. Here, without access to g, they instead respond to the provided explanation e_t. Upon receiving e_t, the agent modifies their covariate from $\ddot{x}_t$ to x_t. As agents' behaviour differs across various explanation types, we define a general reaction model: $x_t := \psi(\ddot{x}_t, e_t, z_t, c_t)$ where ψ is a deterministic and measurable function. After reporting x_t, the agent receives the final prediction $\hat{y}_t := g(x_t)$ and realises the

outcome $y_t := h(x_t, z_t)$, concluding the round. We use P_X to denote the distribution of the random variable X that corresponds to the final covariate x_t.

DM's objective. A typical objective in strategic learning is to minimise the prediction errors while accounting for the shift in agents' features (e.g., (Hardt et al., 2016; Levanon and Rosenfeld, 2021)). Other formulations aim to maximise agents' outcomes or welfare (Xie and Zhang, 2024; Vo et al., 2024). In contrast, our work focuses on the problem of designing explanations and therefore, is agnostic to the DM's objective. Hence, our theoretical results are stated independently of the DM's goal. Section 4 illustrates how a DM can apply our framework through example objectives.

3 AGENTS' RESPONSES TO EXPLANATIONS

Without knowing the predictive model g, the agent has to base their response x_t on the explanation e_t. Hence, the response may be suboptimal, reducing their true utility (in Equation (1)). We formalise the desirable agent's behaviour through the following key concept:

Definition 3.1 (Non-harmful responses)**.** Let $\nu_t = \{x \in \mathcal{X} : u_t(g, x) \geq u_t(g, \ddot{x}_t)\}$. An agent's response, $x_\bullet$, is a non-harmful response if $x_\bullet \in \nu_t$.

Our goal is to characterise explanation methods in terms of their impact on agent behavior, identify conditions under which harmful responses may occur, and develop safeguards to prevent such outcomes. We focus on explanation types that are (i) previously studied in strategic learning, or (ii) accompanied by a clear behavioral model specifying how agents react. In particular, we emphasise actionable explanations that enable agents to improve their prediction outcomes, rather than merely offering post-hoc interpretability. Many popular explanation methods (e.g., feature attribution techniques such as SHAP) lack such actionable guidance (Molnar, 2020), making agent behavior difficult to model. Example A.1 illustrates this issue.

3.1 Surrogate Models

We begin by analysing the use of surrogate models as explanations, e.g., Taylor expansions (Xie and Zhang, 2024), in strategic learning. These provide interpretable approximations of the predictive model g, enabling the agent to construct a surrogate utility function and plan their responses accordingly. In what follows, we derive a necessary condition that ensures such explanations do not induce harmful responses.

Let $f_t : \mathcal{X} \to \mathcal{Y}$ be a surrogate model of g. When the agent observes only f_t, it is natural to assume that they act to maximise the surrogate utility: $u_t(f_t, x) = (-f_t(x)) - c_t(\ddot{x}_t, x)$ by interpreting $-f_t(x)$ as a proxy for benefit instead of $-g(x)$. This assumption is standard when agents lack access to g (see, e.g., (Jagadeesan et al., 2021; Ghalme et al., 2021; Bechavod et al., 2022; Xie and Zhang, 2024)). Hence, their best response becomes $x_t := \arg\max_x u_t(f_t, x)$. Since the surrogate utility $u_t(f_t, \cdot)$ differs from the true utility $u_t(g, \cdot)$, x_t may reduce the agent's true utility. To mitigate this risk, we establish a necessary condition:

Theorem 3.2 (Necessary condition)**.** *Given an agent t with the base covariate $\ddot{x}_t$ who best responds against the surrogate utility function $u_t(f_t, \cdot)$. If it holds, for every possible cost function c_t (Definition 2.2), that the resulting best response x_t belongs to the non-harmful set ν_t (Definition 3.1), i.e., $u_t(g, x_t) \geq u_t(g, \ddot{x}_t)$, then the following also holds:*

$$f_t(\ddot{x}_t) - f_t(x) \leq g(\ddot{x}_t) - g(x) \quad \forall x \in \mathcal{X}_t^{g\downarrow}, \tag{2}$$

with $\mathcal{X}_t^{g\downarrow} := \{x : g(x) < g(\ddot{x}_t)\}$ the set of potential responses with lower scores for the agent.

This theorem says that the surrogate f_t should not overstate the agent's potential gain relative to the true model g. If Equation (2) is violated, there exists a cost function c_t under which the agent is incentivised to choose a response x_t whose cost outweighs the actual gain $g(x_t)$, thereby reducing their true utility. Hence, Equation (2) is a necessary safeguard: *if it fails, harmful responses are possible.* Appendix A.3 gives an example to illustrate how an agent can be misled.

In practice, this safeguard guides the DM in using surrogate-based explanations. It rules out designs that exaggerate agents' perceived gains, a problem common to many popular methods such as LIME (Ribeiro et al., 2016), SHAP (Lundberg, 2017), and Taylor expansions (Xie and Zhang, 2024), which do not satisfy Equation (2) by design (see e.g., Section 4.1).

Moreover, many *noisy* agent reaction models (Rosenfeld et al., 2020; Jagadeesan et al., 2021; Bechavod et al., 2022) can be interpreted as agents responding to some surrogate function f_t. Our necessary condition thus extends to those settings as well, offering guidance on designing communication strategies to ensure agents' non-harmful responses. Similarly, this also extends to broader agent models where agents form beliefs about g based on explanations, a setup that is considered in the recent work by Cohen et al. (2024). Appendix B.2 discusses this extension.

3.2 Action Recommendation-based Explanations

We consider explanations of the form $e_t = (\vec{x}_t, \hat{\vec{y}}_t)$ where $\vec{x}_t \in \mathcal{X}$ denotes a recommended covariate update suggested by the DM and $\hat{\vec{y}}_t := g(\vec{x}_t)$ corresponds

to the predicted outcome if the agent follows this recommendation. We refer to an explanation of this type as an *action recommendation-based explanation* (ARex, plural: ARexes). This class of explanations is desirable because, as we show later, it is sufficient for inducing agents' non-harmful responses.

Formally, an ARex policy is a mapping $\sigma : \mathcal{X} \times \mathcal{G} \to \mathcal{X} \times \mathcal{Y}$. A common design choice for σ is to recommend a minimal feature modification that achieves a desired prediction, often studied as counterfactual explanations (CEs) in explainable ML literature (Molnar, 2020). Throughout, we use ARex as a general term for an explanation of the form $(\vec{x}_t, \hat{\vec{y}}_t)$ regardless of how it is generated. When σ instantiates a design from the CE literature, e.g., Wachter et al. (2017b), we may also call the resulting explanations CEs. This distinction allows us to analyse ARexes broadly without committing to any particular design of σ, such as those that minimise feature modifications (Molnar, 2020) or those that provide *causally plausible* changes in algorithmic recourse (Karimi et al., 2022).

Compared to the previous subsection where the agent has to infer feature updates based on a surrogate model f_t, an ARex explicitly recommends an action $\vec{x}_t$ and reveals its predicted outcome $\hat{\vec{y}}_t$. Thus, the agent no longer needs to infer the predictive model g or speculate about alternative feature changes. Precisely, we follow Tsirtsis and Gomez Rodriguez (2020) and assume that the agent chooses between keeping $\ddot{x}_t$ and adopting $\vec{x}_t$ by comparing their utilities:

$$x_t := \vec{x}_t \;\text{ if }\; u_t\big(g, \vec{x}_t\big) \geq u_t\big(g, \ddot{x}_t\big), \;\text{ else }\; \ddot{x}_t. \tag{3}$$

That is, the agent adopts the recommendation $\vec{x}_t$ if it improves their utility relative to staying with $\ddot{x}_t$. Since the DM discloses both $\hat{\ddot{y}}_t$ and $\hat{\vec{y}}_t$, the agent can evaluate and compare the two utility values directly. This behavioural model assumes that if $\vec{x}$ does not improve the agent's utility, they will keep their base covariate $\ddot{x}$ rather than exploring some other action x'. We view this as a minimal rationality assumption: without reliable knowledge of the predictive model g, agents avoid experimenting with costly alternatives whose benefits are uncertain. This setup lets us ***study the effect of explanations in isolation*** from other side information agents may possess, e.g., from past experience or repeated interactions. Such a focus is both compatible with our research goal and realistic in domains where exploration is costly or restricted, e.g., insurance or credit pricing. Appendix B.3 elaborates on this and discusses alternative behavioral models.

The '$\geq$' sign in Equation (3) lets the agent break ties in favor of the recommendation $\vec{x}_t$, because in many applications, a more favorable prediction $\hat{\vec{y}}$ is linked to better long-term outcomes (e.g., improved financial status, or lower accident risk in insurance pricing). It follows directly that, with ARexes, an agent's best response never harms their true utility (Remark 3.3) and that the class of ARexes suffices to induce all non-harmful responses from agents (Proposition 3.4).

Remark 3.3. For an agent t, any ARex policy σ will induce a best response $x_\bullet$ that belongs to the set of this agent's non-harmful actions ν_t.

Proposition 3.4 (Sufficiency of ARexes)**.** *Consider a population of T heterogeneous agents, indexed by $t \in [T]$ and their base covariates $(\ddot{x}_t)_{t\in[T]}$. Suppose agents respond to ARexes according to Equation (3). Then, for any tuple of (different) explanations $(e_t^\bullet)_{t\in[T]}$ that induce non-harmful responses $(x_t^\bullet)_{t\in[T]}$, there exists a tuple of only ARexes $(e_t^\diamond)_{t\in[T]}$ whose induced responses $(x_t^\diamond)_{t\in[T]}$ satisfy $(x_t^\diamond)_{t\in[T]} = (x_t^\bullet)_{t\in[T]}$.*

However, this alone does not guarantee that an ARex policy in practice can generate such a tuple of ARexes, as agents' observable features $\ddot{x}$ might not be informative enough. Since an ARex induces at most two actions, i.e., Equation (3), it may fail to capture the diversity of responses that other explanations could yield among agents with the same base covariate (see, e.g., Appendix A.4). To that end, we introduce the following condition that specifies how much knowledge the DM must obtain, in $\ddot{x}$, to align agents' responses. This condition *does not restrict agents' heterogeneity* and is formalised as the following assumption.

Assumption 3.5 (Conditional homogeneity of agents' responses)**.** Given an arbitrary explanation method characterised by $(\mathcal{E}, \sigma)$, for any subset of T' agents who share the same base covariate and receive the same explanation, i.e., $(\ddot{x}_t, e_t) = (\ddot{x}, e)$ for all $t \in [T']$, their responses must be identical: $x_t = x_\bullet$ for all $t \in [T']$ and for some $x_\bullet \in \mathcal{X}$.

Although appearing strong at first glance, this assumption is in fact weaker than the standard premise in information design (Bergemann and Morris, 2019) and strategic learning with Bayesian persuasion (Harris et al., 2022a) where a DM is expected to know exactly how agents respond so that their recommendation policy induces obedience, i.e., agents always follow recommendations. Instead, we only require that the DM has "*just enough*" information—captured by $\ddot{x}_t$—such that, conditional on $\ddot{x}_t$, the unobserved factors that directly affects agents' responses are homogeneous. Thus, this places no restriction on agent heterogeneity. Appendix A.5 elaborates on these points and gives an example to illustrate their significance.

In practice, this assumption can be enforced by having the DM query additional information from agents prior to generating explanations. For instance, a car insurer might survey customers on whether they pre-

fer completing a defensive driving course or installing a telematics device. Using this information, the insurer can recommend an action aligned with the customers' preferences, helping them obtain a lower premium.

Theorem 3.6 (Sufficiency of ARex methods)**.** *Suppose conditions in Proposition 3.4 hold. We assume further that Assumption 3.5 holds for all explanation methods. Then, for any arbitrary explanation method* $(\mathcal{E}, \sigma)$ *that induce non-harmful responses* $(x_t^\bullet)_{t\in[T]}$*, there exists an ARex method* $(\mathcal{E}', \sigma')$ *whose induced responses* $(x_t^\diamond)_{t\in[T]}$ *satisfy* $(x_t^\diamond)_{t\in[T]} = (x_t^\bullet)_{t\in[T]}$.

This theorem implies that when evaluating the impact of a DM's predictive model, it suffices to focus solely on ARexes. To identify an optimal explanation method under the no-harm constraint (Definition 3.1), it suffices to search within the space of ARex methods. Consequently, methods outside this class, e.g., LIME, cannot outperform optimal ARexes under the no-harm requirement, regardless of the DM's objective. This sufficiency property is analogous to that of Bayes correlated equilibria (BCE)[3] in information design (Bergemann and Morris, 2019), but with an important distinction: BCE imposes a stronger condition. A BCE can be interpreted as an ARex policy that additionally satisfies an obedience-inducing constraint. Thus, while BCE suffice to rationalise any agent behavior, ARex methods in our setup suffice to rationalise any *non-harmful* agent behavior.

To summarise, ARexes are theoretically desirable because by design, they prevent agents from being misled into harmful actions. Furthermore, under conditional homogeneity, they form a sufficient class of non-harmful explanation methods.

ARexes in practice. While our theory holds regardless of the DM's objective, the DM can also design their ARex policy σ to better serve specific goals. For example, they may jointly optimise both σ and the predictive model g, rather than relying on fixed designs as in standard CE methods (Molnar, 2020). This lets the DM tailor σ to the task at hand. We illustrate this through two concrete scenarios next.

4 EMPIRICAL STUDIES

We empirically evaluate two key properties of ARexes: (i) their no-harm guarantee (Remark 3.3), and (ii) their practical value when optimised for specific objectives. Section 4.1 validates the no-harm guarantee by comparing ARexes against two surrogate-based methods. Sections 4.2 and 4.3 then show how the DM can improve predictive performance by jointly optimising g and σ. Appendix D provides the full setup details.

4.1 On the No-harm Guarantee of ARexes

With a synthetic experiment, we demonstrate that ARexes guarantee no harmful agents' responses whereas Taylor expansions (used as surrogates (Xie and Zhang, 2024)) and LIME (Ribeiro et al., 2016) do not. While LIME is practical and comes with an official Python package, Taylor expansions as explanations (Taylor-ex) provide a clear reaction model, allowing exact computation of agents' responses. We use a quartic function as the predictive model of the DM where $g(x) = x^4 - x^2 + 1$, with 2nd-order Taylor expansions as surrogates. For LIME, we construct a linear approximation of g for each agent. For simplicity, ARexes are generated randomly. We generate 100 agents with a scalar feature $\ddot{x}_t \in \mathbb{R}$ and use the cost function $c_t(\ddot{x}_t, x) = |\alpha_t|\|\ddot{x}_t - x\|_2^2$, with $\alpha_t \in \mathbb{R}$ being the cost factor, reflecting heterogeneity in agents.

Results. Figure 2a plots the change in agents' utility values before and after performing best responses. Taylor-ex and LIME respectively misled 49% and 95% of agents to reducing their utility. Although these two approximate local structures of g, they may exaggerate agents' gains and mislead them into taking costly actions. In contrast, ARexes do not mislead agents, even when the recommendations are generated arbitrarily.

4.2 A Synthetic Experiment

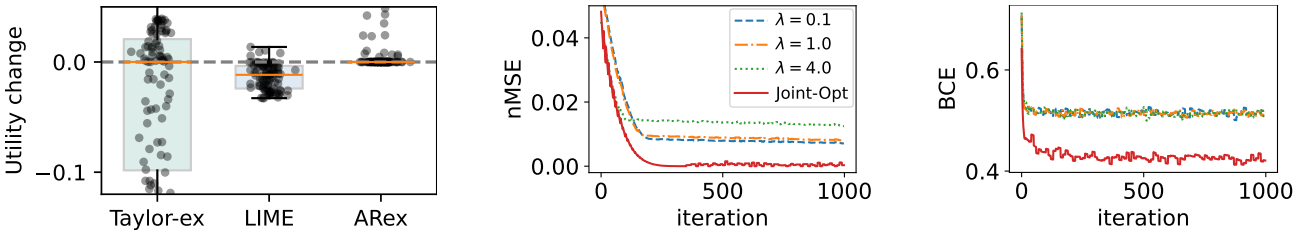
Next, we study a synthetic insurance pricing task that mirrors the setup in Section 2. The DM aims to choose the predictive model g and the ARex policy σ that jointly minimise the expected prediction error. We refer to our approach as *joint-optimisation* (Joint-Opt).[4]

We construct a dataset with agents of a 3-dimensional (observable) feature vector $\ddot{x}_t \in \mathbb{R}^3$ and a scalar (unobservable) feature $z_t \in \mathbb{R}$. Each agent t has the cost function $c_t(\ddot{x}_t, x) = |\alpha_t|\|\ddot{x}_t - x\|_2^2$ where α_t is correlated with $\ddot{x}_t$ and z_t. The outcome function $h : \mathbb{R}^4 \to \mathbb{R}$ is a quadratic function of the concatenated features (x_t, z_t). Here, the DM's goal is to learn g and σ that jointly minimise the expected squared loss: $\min_{g,\sigma} \mathbb{E}_{P_{X,Z}}[(g(X) - h(X, Z))^2]$ where the DM's choice of $\{g, \sigma\}$ affects P_X, the distribution of agents' responses[5]. This objective reflects a typical goal in

[3] This result of Bergemann and Morris (2019) generalises the idea of *straightforward signal* from Bayesian persuasion (Kamenica and Gentzkow, 2011) to the multi-agent environment, such signal is also called Bayesian incentive-compatible (Harris et al., 2022a).

[4] We provide an example algorithm in Appendix D. Our learning procedure extends beyond the classification and discrete setting in Tsirtsis and Gomez Rodriguez (2020), aiming for broader applicability.

[5] For brevity, we omit the explicit connection between X and its optimisation arguments from the objective func-



(a) Agents' utility change. (b) Synthetic data. (c) German credit data.

Figure 2: (a) Taylor-ex and LIME mislead agents into reducing their utility while ARexes do not. The box plot shows the change in agents' utility after best responding. (b) & (c): Joint-Opt has the lowest training-loss curves (nMSE and BCE) against the three baselines, showing that jointly optimising both g and σ is more beneficial to the DM.

insurance pricing: setting premiums that align with future accident costs. Underpricing may lead to financial losses, while overpricing could drive customers away. However, jointly optimising $\{g, \sigma\}$ is challenging due to their interdependence (see Definition 2.1). To address this, observe that for any given pair $\{g, \sigma\}$ that generates ARexes, there exists an equivalent pair $\{g, \sigma^r\}$ that could generate the same, where we call $\sigma^r : \ddot{x}_t \mapsto \vec{x}_t$ an *action recommendation function*. This lets us rewrite the loss as a function of (g, σ^r), yielding:

$$\min_{g,\sigma^r} \ \mathbb{E}_{P_{X,Z}}[(g(X) - h(X, Z))^2]. \tag{4}$$

In practice, because the DM have no access to the outcome function h and the unobserved z_t, an efficient way to solve Equation (4) is through the repeated risk minimisation (RRM) (Perdomo et al., 2020). While training g in RRM is straightforward, the same does not apply to σ^r because each step of risk-minimisation only works if the prediction, i.e., $\hat{y}_t := g(x_t)$, changes whenever σ^r is updated. Thus, the DM must simulate how agents adapt their responses x_t to changes in σ^r. As part of the learning procedure, the DM can deploy some (possibly arbitrary) models $\{g, \sigma\}$ to obtain agents' responses, then learn a model $\hat{\psi} : (\ddot{x}_t, \hat{\ddot{y}}_t, \vec{x}_t, \hat{\vec{y}}_t) \mapsto \hat{x}_t$ that predicts agents' responses, similar to the work of Xie and Zhang (2024). We defer complete details and algorithm to Appendix D.2.

Baselines. We consider counterfactual (CE) policies of the form: $\vec{x}_t := \arg\min_x(g(x) + \lambda\|x - \ddot{x}_t\|_2^2)$ & $\hat{\vec{y}}_t := g(\vec{x}_t)$ and instantiate several values for λ to act as different baselines. Here, g is trained with RRM, but the CE policies remain fixed. No other CE variants are considered, since ARexes subsume CEs and the aim of this experiment is to evaluate joint optimisation of the ARex policy and g, rather than to search for the best CE variant. Since the CE policies are fixed, agents' responses are not simulated here.

Evaluation. We compare the prediction errors on a test set of 10^6 strategic agents. For ease of presentation, the mean-squared errors (MSE) are scaled by a constant and reported as normalised MSE (nMSE): computed on offline data they simply are nMSE, and with agents' strategic responses, *strategic nMSE*.

Results. Figure 2b shows the training loss (nMSE) under RRM and Table 1 reports test performance. Although all methods optimise the predictive model g while accounting for agents' strategic behaviour via RRM, our results show that the choice of explanation policy σ plays a critical role. Specifically, varying λ in fixed CE policies already impacts the predictive performance of g, and jointly optimising σ with g yields further improvement. This highlights the benefit of learning a non-harmful explanation policy tailored to the DM's objective instead of relying on fixed designs.

Although not the main objective, Joint-Opt also achieves full compliance in this synthetic setup. This occurs because the ARex policy σ, through optimisation, learns to *convincingly* guide agents towards regions where g performs well. While this notion of compliance is simulated rather than real human behaviour, it suggests that optimising explanations with strategic dynamics in mind can be effective. Evaluating this effect in user studies or behavioural experiments is a promising direction for future work.

4.3 German Credit Dataset

For the German credit dataset (Hofmann, 1994) with a classification task, we follow Xie and Zhang (2024) to preprocess the data and to simulate strategic behaviour. Next, a logistic regression model is fitted on the original data to estimate y_t, and used to simulate outcomes when agents modify their covariates x_t. We use CTGAN (Xu et al., 2019) to generate 9,000 more samples for training and 1,000 for testing.

DM-agents interactions. The DM's predictive model is a binary classifier $g(x) := \mathbb{1}[g_s(x) \geq 0.5]$ where the underlying scoring $g_s(x) \in [0, 1]$ outputs the predicted probability of a positive outcome. Agents aim to maximise $g_s(x)$, yielding utility

tion, though a full expansion is provided in Appendix C.1.

Table 1: Joint-Opt achieves the lowest test error on synthetic data and highest test score on real-world data, while maintaining strong compliance rates, which indicates the portions of agents that follow the DM's recommended actions.

		Joint-Opt	$\lambda = 0.1$	$\lambda = 1.0$	$\lambda = 4.0$
Synthetic data	Strategic nMSE (↓)	**2e-4**	7e-3	8e-3	1e-2
	Compliance rate	1.0	1.0	1.0	1.0
German credit data	Strategic F_1 score (↑)	**0.9**	0.86	0.84	0.85
	Compliance rate	0.84	1.0	1.0	1.0

$u_t(g_s, x)$. We design the cost function as $c(\ddot{x}_t, x) := 0.01 \sum_{i \in \mathcal{I}} |\ddot{x}_{ti} - x_i| / (x_i^U - x_i^L)$, where $\mathcal{I}$ is the set of modifiable features and $[x_i^L, x_i^U]$ denotes the valid range of feature i.

Evaluation. Counterfactual explanations are generated similarly to the synthetic setup, with additional constraints (e.g., categorical/bounded features) enforced via projected gradient descent. Unlike previous setup, we now use $-g_s(x)$ in the objective function for generating CEs as the agents now benefit from higher prediction scores. We evaluate predictive accuracy, with F_1 score, on the test set of 1,000 strategic agents, referring to this as *strategic* F_1 *score* to reflect shifts in (x_t, y_t) due to strategic behaviour.

Results. Figure 2c shows the training loss, i.e., binary cross entropy (BCE), under RRM, and Table 1 reports test performance. These results support our findings in the previous synthetic setup: joint optimisation of g and σ yields the most favourable outcome for the DM, while maintaining a reasonably high compliance rate under simulated strategic behaviour.

Finally, while our theory focuses on tabular data, Appendix D.4 demonstrates its application to images in a more complex setting.

5 RELATED WORK

In this section, we overview related work and provide complete details in Appendix E.

Strategic learning. Several studies have examined how partial information of the DM's predictive model shapes agents' decisions e.g., Jagadeesan et al. (2021); Ghalme et al. (2021); Bechavod et al. (2022); Haghtalab et al. (2024); Xie and Zhang (2024). In particular, the works of Harris et al. (2022a) and Cohen et al. (2024) are closest to ours. As discussed in Section 3.2, Harris et al. (2022a) focuses on the obedience-inducing property (Bayesian incentive compatibility) of a subclass of action recommendations, whereas we focus on the no-harm property of action recommendations. Unlike action recommendations, Cohen et al. (2024) instead releases a subset of the hypothesis class to all agents, aligning with global explanations in our framework (Section 2). However, interpreting a set of models—such as neural networks—can be difficult for agents. In contrast, ARexes are both more interpretable and guaranteed not to mislead agents.

Counterfactual explanations (CEs) and algorithmic recourse. Tsirtsis and Gomez Rodriguez (2020); Karimi et al. (2022) explore CEs and algorithmic recourse, for strategic agents. Although algorithmic recourse focuses on recommending actions to achieve better outcomes, its implementation often relies on strong causal assumptions, which can be impractical when such knowledge is unavailable. In contrast, our work adopts a weaker notion of desirability centred on agents' welfare—ensuring non-harmful responses—and examines a broader range of explanation types beyond counterfactuals. Although both Tsirtsis and Gomez Rodriguez (2020) and our work involve CEs, the contributions differ. Precisely, they focus on optimising CEs in strategic settings, while we analyse multiple explanation types and formally show why ARexes are more desirable. In addition, our proposed learning procedure in Section 4.2, though not the main focus, is designed to be more general, extending beyond the classification and discrete case in Tsirtsis and Gomez Rodriguez (2020).

Information design. The extensive literature on information design, surveyed by Bergemann and Morris (2019), studies how to design information disclosure policies in a game of two parties. While our results are inspired by these works, e.g., Theorem 3.6, the goals differ significantly. As discussed in Section 3.2, information design aims at *persuading* agents with a general response model and does not necessarily ensure the no-harm property (Definition 3.1). In contrast, we study explanation methods that prioritise the no-harm property, safeguarding agents' welfare. By incorporating specific agent models in strategic settings, we establish the sufficiency of ARexes without requiring the DM to know exactly how agents respond.

6 CONCLUSION

To summarise, we address the challenge of providing safe, actionable explanations in strategic learning, where DMs must balance transparency with utility optimisation. We formalise action recommendation-

based explanations (ARexes), which ensure that agents act without incurring self-harm. By introducing the no-harm property and a conditional homogeneity assumption, we demonstrate that ARexes let DMs achieve optimal outcomes while safeguarding agent welfare. Consequently, our work clarifies the distinctions of different explanation types through the lens of strategic learning. Finally, we propose a framework to jointly optimise predictive models and ARex policies, aligning the DM's objectives with agents' responses.

Our findings rest on commonly adopted assumptions about agents' behavior, such as their utility functions or reaction models. While these assumptions may limit the generalisability of our approach, they do not diminish its broader relevance. Intuitively, when explanations omit certain information, conditions are needed to ensure agents' inferred gains are not exaggerated, protecting their welfare. ARexes succeed in this regard by being transparent about the potential benefits of recommended actions, eliminating the risk of misleading agents. Future work could explore diverse agent behavior models, dynamic environments, and more scalable learning algorithms to enhance the applicability and efficiency of this approach.

Acknowledgements

We sincerely thank the members of the Rational Intelligence (RI) Lab, including Abbavaram Gowtham Reddy, Anurag Singh, Monseej Purkayastha, and Swathi Suhas, for their insightful discussions, constructive feedback, and invaluable contributions to this work. We also extend our gratitude to the visiting researchers, Amin Charusaie, Majid Mohammadi, Masaki Adachi, Rattaya Kaewvichai, and Saptarshi Saha, for their stimulating discussions and fresh perspectives, which enriched our understanding of the problem. Their contributions have been greatly appreciated.

We also thank the anonymous reviewers for their valuable feedback to improve our work.

Yixin Wang was supported in part by funding from the Office of Naval Research under grant N00014-23-1-2590, the National Science Foundation under grant No. 2310831, No. 2428059, No. 2435696, No. 2440954, a Michigan Institute for Data Science Propelling Original Data Science (PODS) grant, Two Sigma Investments LP, and LG Management Development Institute AI Research.

Kiet Q. H. Vo is a doctoral candidate at Saarland University.

References

Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 6–25, 2021.

Saba Ahmadi, Avrim Blum, and Kunhe Yang. Fundamental bounds on online strategic classification. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 22–58, 2023.

Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *International Conference on Machine Learning*, pages 1691–1715. PMLR, 2022.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

Siu Lun Chau, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. Rkhs-shap: Shapley values for kernel methods. *Advances in neural information processing systems*, 35:13050–13063, 2022.

Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. *Advances in Neural Information Processing Systems*, 36:50769–50795, 2023.

Molnar Christoph. *Interpretable machine learning: A guide for making black box models explainable*. Leanpub, 2020.

Lee Cohen, Saeed Sharifi-Malvajerdi, Kevin Stangl, Ali Vakilian, and Juba Ziani. Sequential strategic screening. In *International Conference on Machine Learning*, pages 6279–6295. PMLR, 2023.

Lee Cohen, Saeed Sharifi-Malvajerdi, Kevin Stangl, Ali Vakilian, and Juba Ziani. Bayesian strategic classification. *arXiv preprint arXiv:2402.08758*, 2024.

Council of European Union. Council regulation (EU) no 679/2016, 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classi-

fication in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36, 2024.

Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 2018.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Keegan Harris, Hoda Heidari, and Steven Z Wu. Stateful strategic regression. *Advances in Neural Information Processing Systems*, 34:28728–28741, 2021.

Keegan Harris, Valerie Chen, Joon Kim, Ameet Talwalkar, Hoda Heidari, and Steven Z Wu. Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35:11131–11144, 2022a.

Keegan Harris, Dung Daniel T Ngo, Logan Stapleton, Hoda Heidari, and Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pages 8502–8522. PMLR, 2022b.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. *CoRR*, abs/1808.08646, 2018. URL http://arxiv.org/abs/1808.08646.

Jessica Hullman, Ziyang Guo, and Berk Ustun. Explanations are a means to an end. *arXiv preprint arXiv:2506.22740*, 2025.

Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, pages 4687–4697. PMLR, 2021.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

Amir-Hossein Karimi, Krikamol Muandet, Simon Kornblith, Bernhard Schölkopf, and Been Kim. On the relationship between explanation and prediction: A causal view. In *International Conference on Machine Learning*, pages 15861–15883. PMLR, 2023.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pages 6243–6253. PMLR, 2021.

Charles Lu, Baihe Huang, Sai Praneeth Karimireddy, Praneeth Vepakomma, Michael Jordan, and Ramesh Raskar. Data acquisition via experimental design for data markets. *Advances in Neural Information Processing Systems*, 37:118086–118118, 2024.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on*

knowledge discovery and data mining, pages 1135–1144, 2016.

Nir Rosenfeld, Anna Hilgard, Sai Srivatsa Ravindranath, and David C Parkes. From predictions to decisions: Using lookahead regularization. *Advances in Neural Information Processing Systems*, 33:4115–4126, 2020.

Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017.

Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *International Conference on Machine Learning*, pages 8676–8686. PMLR, 2020.

Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. Pac-learning for strategic classification. *Journal of Machine Learning Research*, 24(192):1–38, 2023.

Stratis Tsirtsis and Manuel Gomez Rodriguez. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems*, 33:16749–16760, 2020.

Kiet QH Vo, Muneeb Aadil, Siu Lun Chau, and Krikamol Muandet. Causal strategic learning with competitive selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15411–15419, 2024.

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law*, 7(2):76–99, 2017a.

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

Tian Xie and Xueru Zhang. Non-linear welfare-aware strategic learning. *arXiv preprint arXiv:2405.01810*, 2024.

Tian Xie, Zhiqun Zuo, Mohammad Mahdi Khalili, and Xueru Zhang. Learning under imitative strategic behavior with unforeseeable outcomes. *arXiv preprint arXiv:2405.01797*, 2024.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Explanation Design in Strategic Learning: Sufficient Explanations that Induce Non-harmful Responses Supplementary Materials

A ADDITIONAL ILLUSTRATIVE EXAMPLES

This section contains more examples to illustrate our theory.

A.1 Examples of local and global explanations

In this work, we adopt a broad notion of explanations consistent with prior literature, where an explanation is defined as an element of an explanation space produced by a mapping σ, and we place no restriction on its epistemic usefulness (i.e., its ability to support model understanding). In particular, some explanation formats—such as counterfactual explanations or partial model descriptions—may offer limited insight into the internal workings of the predictive model, even though they are widely treated as explanations. This perspective aligns with formal frameworks that model explanations as abstract objects without imposing epistemic constraints (Karimi et al., 2023; Hullman et al., 2025). Our goal is not to characterise or evaluate explanations based on their epistemic usefulness, but rather to analyse how different forms of information communicated by the DM influence agents' behaviour. We refer to Appendix B.1 for a discussion of the distinction between epistemic and strategic roles of explanations.

We provide some specific examples for global and local explanations that fit into our setting (Definition 2.1):

- Global surrogate models such as linear models or decision trees that approximate the DM's predictive model g (Molnar, 2020). In this scenario, a *constant* explanation $e \in \mathcal{E}$, regardless of $\ddot{x}_t$, is some surrogate model $f : \mathcal{X} \to \mathcal{Y}$ for the true model $g : \mathcal{X} \to \mathcal{Y}$. The explanation space $\mathcal{E}$ is a subset of linear models or decision tree models: $\mathcal{F} = \{f' : \mathcal{X} \to \mathcal{Y}\}$.
- Partial descriptions of the DM's predictive model (Cohen et al., 2024). In this scenario, a *constant* explanation $e \in \mathcal{E}$, regardless of $\ddot{x}_t$, is a *partial* description (of g) that corresponds to some subset of the hypothesis space $\mathcal{G}_S \subseteq \mathcal{G}$ such that $g \in \mathcal{G}_S$. This partial description narrows down the agents' uncertainty about g without fully revealing g.
- Feature attribution-based explanation methods that assign importance scores to features such as SHAP (Lundberg, 2017). For example, when the covariate $\ddot{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ is a vector of d features, an explanation $e_t = (e_{t1}, \ldots, e_{td}) \in \mathcal{E} \subseteq \mathbb{R}^d$ is a vector containing the importance scores of each features in $\ddot{x}_t$.
- Local surrogate models such as Taylor expansions (Xie and Zhang, 2024). An explanation e_t is some function $f_t : \mathcal{X} \to \mathcal{Y}$ that approximates g in a local neighbourhood of $\ddot{x}_t$ and the explanation space $\mathcal{E}$ is a subset of all such functions, e.g., $\mathcal{E} \subseteq \mathcal{F} = \{f' \mid f : \mathcal{X} \to \mathcal{Y}\}$.
- Counterfactual explanations (Wachter et al., 2017b). In this scenario, $e_t = (\vec{x}, g(\vec{x}))$ and $\mathcal{E} = \{(\vec{x}, g(\vec{x})) : g(\vec{x}) < \hat{\hat{y}}\}$, where $\vec{x}$ denotes the recommended covariate for the agent to change to, in order to receive a more favourable prediction, i.e., $g(\vec{x})$ from the DM, e.g., lower insurance premium: $g(\vec{x}) < \hat{\hat{y}}$.

A.2 A Shapley value example

We provide a simple example using Shapley values to illustrate the point made in Section 3: many attribution methods fail to offer clear, actionable guidance for strategic agents. This limitation arises because Shapley values depend on the underlying data distribution and therefore may not reliably capture the behavior of the DM's predictive model g.

Example A.1. Consider the predictive model $g(x) = x_1 - x_2^2$ for any $x = [x_1 \quad x_2]^\top \in \mathbb{R}^2$, and an agent with the base feature vector $\ddot{x} = [16 \quad 4]^\top$. With a single agent, we drop the subscript t for simplicity. Furthermore, suppose that the 2 features $\ddot{X}_1, \ddot{X}_2$ are statistically independent and that $\ddot{X}_2 \sim \mathcal{U}([2,5])$. The Shapley value of this agent's 2nd feature is

$$\begin{aligned}
&\phi_2(g, \ddot{x}) \\
&= \frac{1}{2}\Big(g(\ddot{x}) - \mathbb{E}\left[g(\ddot{X}_1 = 16, \ddot{X}_2)\right] \\
&\qquad + \mathbb{E}\left[g(\ddot{X}_1, \ddot{X}_2 = 4)\right] - \mathbb{E}\left[g(\ddot{X}_1, \ddot{X}_2)\right]\Big) \\
&= \frac{1}{2}\Big(0 - \mathbb{E}\left[16 - \ddot{X}_2^2\right] + \mathbb{E}\left[\ddot{X}_1 - 16\right] - \mathbb{E}\left[\ddot{X}_1 - \ddot{X}_2^2\right]\Big) \\
&= \frac{1}{2}\Big(-32 + 2\mathbb{E}\left[\ddot{X}_2^2\right]\Big) = -3 < 0.
\end{aligned}$$

However, if we consider $\ddot{X}_2 \sim \mathcal{U}([2,8])$, then $\mathbb{E}\left[\ddot{X}_2^2\right] = 28$ and the Shapley value for this 2nd feature is

$$\phi_2(g, \ddot{x}) = \frac{1}{2}\Big(-32 + 2\mathbb{E}\left[\ddot{X}_2^2\right]\Big) = 12 > 0.$$

On the other hand, the partial derivative of the function g at the point $x_2 = \ddot{x}_2 = 4$ is

$$\frac{dg}{dx_2}(x_2 = 4) = -8,$$

which implies that this agent can achieve a lower prediction score by increasing the value of their 2nd feature $\ddot{x}_2$. However, the sign of the Shapley value for this feature, as we have shown, can vary depending on the distribution of the data, as a result, this Shapley value cannot say how this agent should change their feature to obtain better prediction score.

A.3 A misled agent

We present a simple example showing how linear surrogate models cannot guarantee that the induced agent's responses are non-harmful (Definition 3.1). This is because they do not satisfy the necessary condition in Theorem 3.2.

Example A.2. Suppose that an insurance company uses the $g(x) = x^2$ to predict the risk of a customer whose feature takes on the base value $\ddot{x} = 5$, which corresponds to the prediction $\hat{\ddot{y}} = 25$. With this single-agent scenario, we drop the subscript t for simplicity. Suppose further that this customer has the following cost function for changing their feature:

$$c(\ddot{x}, x) := c(\Delta x) := \begin{cases} 3(\Delta x)^2 & \forall \Delta x \in (-\infty, -3] \\ 9|\Delta x| & \forall \Delta x \in [-3, 0] \\ > 0 & \forall \Delta x \in (0, \infty), \end{cases}$$

where we use Δx to denote $x - \ddot{x}$ for any $x \in \mathcal{X}$, and rewrite the cost function into $c(\Delta x)$ for simplicity. The DM discloses a localised linear model $f(x) = 10x - 25$ tangent to $g(x)$ at the base value $\ddot{x} = 5$. As the agent wants to minimise their predictive risk, their true utility function and surrogate utility function, to be maximised are respectively

$$\begin{aligned}
u(g, x) &= -g(\ddot{x} + \Delta x) - c(\Delta x) = -(5 + \Delta x)^2 - c(\Delta x), \\
u(f, x) &= -f(\ddot{x} + \Delta x) - c(\Delta x) = -10(5 + \Delta x) + 25 - c(\Delta x).
\end{aligned}$$

Hence, $x = 2$ (or equivalently $\Delta x = -3$) is the customer's best response as it maximises their surrogate utility function $u(f, x)$. However, this leads to a reduction in the true utility function, i.e., $u(g, 2) < u(g, 5)$, because the customer has paid a high cost only to achieve a small reduction in their prediction value.

A.4 On diversity of agents' responses

We give an example on how ARexes can fail to capture the diversity of the agents' responses that can be induced by other explanation methods. This helps illustrate why it is important for the DM to acquire additional information about agents as implied by Assumption 3.5.

Example A.3. Suppose that there are three agents who have the same base covariates $\ddot{x}$ and that their cost factors are respectively $\alpha_1, \alpha_2, \alpha_3$. Their cost functions have the form $\alpha_t \|x - \ddot{x}_t\|_2^2$. From Definition 2.1, because all these three agents have the same base covariates $\ddot{x}$, they will receive the same explanation, regardless of the explanation method.

Suppose that these agents receive the same surrogate function $f : \mathcal{X} \to \mathcal{Y}$ as an explanation. Then, from Section 3.1, these agents responses are

$$x_t := \arg\max_x \{-f(x) - \alpha_t \|x - \ddot{x}\|_2^2\}.$$

Because the agents' utility functions only differ in α_t, they can result in different maximisers x_t, depending on the choice for surrogate function f. Thus, this surrogate function induces maximally three actions from these agents.

In contrast, if all these agents receive an ARex $(\vec{x}_\bullet, \hat{\vec{y}}_\bullet)$, at most two distinct actions can be induced, as implied by Equation (3).

A.5 On Assumption 3.5

Assumption 3.5 can be viewed as requiring the DM to have enough information about agents (e.g., cost-related information c_t being captured in $\ddot{x}_t$) so that they can partition agents into subgroups with homogeneous responses. It is important to note that this does not assume away the heterogeneity of agents, even within each subgroup.

The following example illustrates the distinction between Assumption 3.5 and the full-information premise in information design, as well as the distinction between our sufficiency result and the concept of obedience in prior work. This also highlight the significance of our assumption as a relaxation of the full-information premise.

Example A.4. Suppose that there is a population of heterogeneous agents (e.g., loan applicants) each of which has observable features x (e.g., income, employment status, credit history) and a hidden continuous feature z (e.g., capturing their financial risk preferences or cost sensitivity). Suppose an agent's response is determined by $\psi(x, \delta(z), e)$ where e is an explanation and $\delta(z)$ bins z into behavioural categories(e.g., risk-averse, risk-neutral, risk-seeking). Suppose there exists a deterministic mapping[6] from x to $\delta(z)$, reflecting the fact that applicants with similar observable features tend to fall into the same behavioral category.

In our work, if the DM observes x and conditions on this, then agents are split into subgroups of homogeneous responses, because of the existence of said deterministic mapping. Thus, Assumption 3.5 holds. However, in information design where obedience is studied, the DM would need to additionally know ψ and $\delta(z)$ to know exactly how an agent responds. This shows the significance of our relaxation: it does not require the DM to know more about agents. Moreover, agents' heterogeneity remains, because within each group, z varies and across the population, both z and $\delta(z)$) vary.

This relaxation has practical implications. Since the DM only needs partial information to construct behaviorally homogeneous groups, lightweight elicitation methods (e.g., surveys or recorded past behavior) may suffice. This makes Assumption 3.5 more applicable than full-information premises.

B ADDITIONAL RESULTS AND EXTENSIONS

B.1 On the epistemic and strategic roles of explanations

Explanations often serve two distinct roles: an epistemic role (helping users understand the model) and a strategic role (guiding users toward beneficial actions). A good explanation should therefore satisfy both roles. However, an explanation may be epistemically sound (i.e., faithful to the underlying prediction model) yet still

[6] Note that this does not necessarily mean that one of them is the cause of the other.

fail strategically if it leads users to take inappropriate actions. More details and illustrative examples can be found in Theorem 3.2 and Appendices A.2 and A.3. When this occurs (e.g., when an explanation inadvertently encourages a costly but unhelpful change), users may experience worse outcomes and view the system as unreliable or unhelpful, ultimately eroding trust in the DM's system. As noted in the Introduction, such effects matter to the decision maker (DM) because user trust is tied to long-term objectives, such as maintaining customer retention or reducing disputes.

B.2 A necessary condition in broader settings

The result in Theorem 3.2 also extends to broader agent models where agents form beliefs about g based on explanations, a setup that is considered in the concurrent work by Cohen et al. (2024).

Corollary B.1 (Necessary condition)**.** *For any agent with a tuple $(\ddot{x}_t, c_t, z_t)$, suppose that (1) $\theta_t \in \Theta \subset \mathbb{R}^d$ is a random variable distributed according to the agent's prior $p(\theta_t)$ over the unknown parameter θ_0 of the true predictive model g_{θ_0}, (2) $p(\theta_t|e_t) \propto p(e_t|\theta_t)p(\theta_t)$ is the posterior belief of this agent upon receiving the explanation e_t, (3) f_t represents the agent's updated belief about g_{θ_0} defined as $f_t(x) := \mathbb{E}_{\theta_t}[g_{\theta_t}(x) \mid e_t], \forall x \in \mathcal{X}$. Then, the result in Theorem 3.2 extends to this setting.*

For Bayesian agents, this corollary shows that the safeguard of Theorem 3.2, i.e., Equation (2), remains the necessary requirement to guarantee non-harmful responses. It thus guides the DM in designing explanations that prevent agents' self-harming actions, while also serving as a benchmark for whether a no-harm guarantee is even achievable. Although this safeguard may be stringent, this reflects its role as a necessary requirement: *if it fails, no explanation policy can ensure that all agents avoid harmful actions.*

B.3 On the behavioural model in response to ARex, in Equation (3)

The behavioural model in Equation (3) which we adopt from prior work (Tsirtsis and Gomez Rodriguez, 2020) assumes that the agent t chooses to either keep their base covariate $\ddot{x}_t$ or follow the recommendation $\vec{x}_t$, rather than exploring some other action x'. Such behaviour implies that the agent lacks side information about the DM's predictive model g and this discourages the agent from picking another covariate update x' whose benefit is unknown. This is, in fact, compatible with the setting we focus on: ***agents rely solely on the explanation provided by the decision maker***. Our goal is then to study how different explanation types, when they are the only source of information, influence agent behavior. In contrast, when agents have access to additional side information that strongly influences their decisions, it becomes difficult to disentangle the behavioural effects of different explanation types. Such settings, where explanations interact with richer sources of information, are an interesting direction for future work.

In practice, there are cases where agents lack side information because repeated exploration might be costly, risky, or restricted. For instance, in insurance pricing repeated probing may be limited to prevent model leakage or may require expensive external data checks. In loan pricing, each application may trigger hard credit checks and fees, discouraging exploration.

Nevertheless, we briefly discuss some alternative models and their implications below.

Multiple recommendations as an ARex. Consider the case where the DM releases an ARex that contains k recommended covariate updates and their respective prediction scores, i.e., $e = \{(\vec{x}, \hat{\vec{y}})_j\}_{j=1}^k$. Assuming that the agent either picks a covariate update out of this recommended set or does nothing, we then define the set of feasible covariate updates for this agent as $\mathcal{X}_t^{\text{fe}} = \{\ddot{x}_t\} \cup \{\vec{x}_j\}_{j=1}^k$. Formally, the behavioural model in Equation (3) becomes:

$$x_t := \arg\max_{x \in \mathcal{X}_t^{\text{fe}}} u_t(g, x).$$

Then, the results in Remark 3.3 and Proposition 3.4 also hold, under some technical assumptions on the predictive model g. In addition, when Assumption 3.5 holds, Theorem 3.6 also holds.

Action exploration with some probability. Consider the following agents' reaction model under ARexes:

$$x_t := \vec{x}_t \text{ if } u_t(g, \vec{x}_t) > u_t(g, \ddot{x}_t)$$
$$\text{else } \begin{cases} \ddot{x}_t \text{ with probability } p, \\ x' \text{ with probability } 1-p, \end{cases}$$

where the choice of an agent's exploration x' depends on both the explanation e_t and any side information that is available to the agent. The agent's choice may reflect prior experience, social influence, or individual behavioural traits (e.g., risk aversion or risk seeking).

Deriving formal guarantees on the agent's behaviour in such settings typically requires strong assumptions about the decision maker's knowledge of the agent's behavioural model, as in information design (Bergemann and Morris, 2019). By contrast, our formulation adopts a simpler behavioural model that avoids such assumptions: we only require the decision maker to partition agents correctly, as we also discuss in Appendix A.5.

C PROOFS OF THE MAIN RESULTS

This section contains the derivations and proofs of our main results presented in Section 3 and Section 4.

C.1 DM's objective

We expand the DM's original objective in Section 4 to show how all the parameters affect the objective:

$$\begin{aligned} &\min_{g,\sigma} \mathbb{E}_{P_{X,Z}} \left[\big(g(X) - h(X,Z)\big)^2 \right] \\ =&\min_{g,\sigma} \mathbb{E}_{P_{\ddot{X},C,Z}} \left[\Big(g\big(\underbrace{\psi(\ddot{X}, \sigma(\ddot{X}, g), Z, C)}_{X} \big) - h\big(\underbrace{\psi(\ddot{X}, \sigma(\ddot{X}, g), Z, C)}_{X}, Z \big) \Big)^2 \right], \end{aligned}$$

where the random variable $X := \psi(\ddot{X}, \sigma(\ddot{X}, g), Z, C)$ denotes the response of an agent and ψ is the response function, as defined in Section 2.

C.2 Proof for Theorem 3.2

We introduce the following lemma to help proving Theorem 3.2.

Lemma C.1. *If the necessary condition (i.e., Equation (2)) is violated, then there exist a value $x_\bullet \in \mathcal{X}_t^{g\downarrow}$ and a cost function c_t satisfying the following three conditions:*

$$\begin{cases} 0 < g(\ddot{x}_t) - g(x_\bullet) < c_t(\ddot{x}_t, x_\bullet), \\ c_t(\ddot{x}_t, x_\bullet) < f_t(\ddot{x}_t) - f_t(x_\bullet), \\ c_t(\ddot{x}_t, x_\bullet) + \Big(f_t(x_\bullet) - f_t(x) \Big) < c_t(\ddot{x}_t, x) \quad \forall x \in \mathcal{X}_t^{g\downarrow} \setminus \{x_\bullet\}. \end{cases} \tag{5}$$

Proof. The first line of Equation (5) says that the change in the prediction value is smaller than the cost for updating the agent's covariate. We have $0 < g(\ddot{x}_t) - g(x_\bullet)$ because of the definition of $\mathcal{X}_t^{g\downarrow}$, and $c_t(\ddot{x}_t, x_\bullet) > 0$ because of Definition 2.2. Because these two definitions are unrelated, there exist infinitely many such pairs of $\{x_\bullet, c_t\}$.

The second line of Equation (5) says that the cost for updating the agent's covariate is smaller than the change in the *surrogate prediction* value (i.e., $f_t(\cdot)$). Because Equation (2) is violated, there exists $x_\bullet \in \mathcal{X}_t^{g\downarrow}$ such that $g(\ddot{x}_t) - g(x_\bullet) < f_t(\ddot{x}_t) - f_t(x_\bullet)$. Given such $x_\bullet$, there exists a cost function c_t such that $g(\ddot{x}_t) - g(x_\bullet) < c_t(\ddot{x}_t, x_\bullet) < f_t(\ddot{x}_t) - f_t(x_\bullet)$. This is because the cost function c_t can be designed independent of $\{g, f_t, \ddot{x}_t, x_\bullet\}$.

Before explaining the meaning of the third line of Equation (5), we show how a pair $\{x_\bullet, c_t\}$ can satisfy this condition. Given $\{x_\bullet, f_t\}$, one can design a cost function c_t that satisfies the following, without violating the

first two conditions of Equation (5):

$$\begin{cases} c_t(\ddot{x}_t, \ddot{x}_t) := 0, \\ c_t(\ddot{x}_t, x) := c_t(\ddot{x}_t, x_\bullet) + \Big(f_t(x_\bullet) - f_t(x)\Big) + \Big|f_t(x_\bullet) - f_t(x)\Big| \quad \forall x \in \mathcal{X}_t^{g\downarrow} \setminus \{x_\bullet\}, \end{cases}$$

where $c_t(\ddot{x}_t, x_\bullet) > 0$ and $\Big(f_t(x_\bullet) - f_t(x)\Big) + \Big|f_t(x_\bullet) - f_t(x)\Big| \geq 0$ for all $x \in \mathcal{X}_t^{g\downarrow} \setminus \{x_\bullet\}$. This holds because we impose no additional restrictions on c_t—such as smoothness—beyond those specified in Definition 2.2. As a result, the design for $c_t(\ddot{x}_t, x)$, for all $x \neq x_\bullet$, is not affected by the design for $c_t(\ddot{x}_t, x_\bullet)$.

This concludes the proof for this lemma. We explain the third condition in Equation (5) in the next proof. □

We now prove Theorem 3.2.

Proof of Theorem 3.2. We prove this by contrapositive. Suppose that the necessary condition (Equation (2)) is violated, then there exist a value $x_\bullet \in \mathcal{X}_t^{g\downarrow}$ and a cost function c_t satisfying Equation (5).

We explain the meaning of the third condition in Equation (5). Observe that this condition is equivalent to

$$\underbrace{-f_t(x_\bullet) - c_t(\ddot{x}_t, x_\bullet)}_{u_t(f_t, x_\bullet)} > \underbrace{-f_t(x) - c_t(\ddot{x}_t, x)}_{u_t(f_t, x)} \quad \forall x \in \mathcal{X}_t^{g\downarrow} \setminus \{x_\bullet\}.$$

Then, using Equation (5), we can see that, if the best response of this agent t, against the local surrogate function f_t, lies in the set $\mathcal{X}_t^{g\downarrow} \cup \{\ddot{x}_t\}$, then $x_\bullet$ is the solution, since

$$x_\bullet := \arg \min_{x \in \mathcal{X}_t^{g\downarrow} \cup \{\ddot{x}_t\}} f_t(x) + c_t(\ddot{x}_t, x) \tag{6}$$

$$:= \arg \max_{x \in \mathcal{X}_t^{g\downarrow} \cup \{\ddot{x}_t\}} u_t(f_t, x). \tag{7}$$

Moreover, because $g(\ddot{x}_t) - g(x_\bullet) < c_t(\ddot{x}_t, x_\bullet)$, as specified in the first condition of Equation (5), we have

$$-\Big(g(x_\bullet) + c_t(\ddot{x}_t, x_\bullet)\Big) < -g(\ddot{x}_t) \tag{8}$$

$$\Rightarrow u_t(g, x_\bullet) < u_t(g, \ddot{x}_t). \tag{9}$$

This means that $x_\bullet \notin \nu_t$ (Definition 3.1).

On the other hand, if the best response is some $x_\square \notin (\mathcal{X}_t^{g^-} \cup \{\ddot{x}_t\})$, we have

$$g(x_\square) + \underbrace{c_t(\ddot{x}_t, x_\square)}_{>0} > g(x_\square) \geq g(\ddot{x}_t), \tag{10}$$

which results in $u_t(g, x_\square) < u_t(g, \ddot{x}_t)$.

In either of both cases ($x_\bullet$ or $x_\square$), the agent's response does not belong to the non-harmful set ν_t. Hence, by the contrapositive, ensuring that the agent's responses lie in ν_t requires that Equation (2) holds. This concludes the proof. □

C.3 A sufficient condition to guarantee non-harmful responses

Theorem C.2 (Sufficient condition)**.** *Given a base covariate value $\ddot{x}_t \in \mathcal{X}$ and a surrogate model $f_t : \mathcal{X} \to \mathcal{Y}$, if it holds that*

$$f_t(\ddot{x}_t) - f_t(x) \leq g(\ddot{x}_t) - g(x) \quad \forall x \in \mathcal{X},$$

then, for any agent with the same base covariate $\ddot{x}_t$, their response $x_\bullet$, against the surrogate utility $u_t(f_t, \cdot)$, lies within their non-harmful set ν_t.

Proof. For any arbitrary agent t with the base covariate $\ddot{x}_t$ and the cost function c_t, suppose we have

$$f_t(\ddot{x}_t) - f_t(x) \leq g(\ddot{x}_t) - g(x) \quad \forall x \in \mathcal{X}. \tag{11}$$

Let $x_\diamond \in \mathcal{X}$ denotes the agent's best response against the surrogate utility $u_t(f_t, \cdot)$, then

$$\begin{aligned} & u_t(f_t, x_\diamond) && \geq u_t(f_t, \ddot{x}_t) \\ & -f_t(x_\diamond) - c_t(\ddot{x}_t, x_\diamond) && \geq -f_t(\ddot{x}_t) - \underbrace{c_t(\ddot{x}_t, \ddot{x}_t)}_{=0} \\ \Rightarrow \quad & f_t(\ddot{x}_t) - f_t(x_\diamond) && \geq c_t(\ddot{x}_t, x_\diamond). \end{aligned}$$

Using the assumed condition in Equation (11), we obtain:

$$\begin{aligned} & c_t(\ddot{x}_t, x_\diamond) && \leq f_t(\ddot{x}_t) - f_t(x_\diamond) \leq g(\ddot{x}_t) - g(x_\diamond) \\ \Rightarrow \quad & c_t(\ddot{x}_t, x_\diamond) && \leq g(\ddot{x}_t) - g(x_\diamond) \\ \Rightarrow \quad & -g(\ddot{x}_t) - \underbrace{c_t(\ddot{x}_t, \ddot{x}_t)}_{=0} && \leq -g(x_\diamond) - c_t(\ddot{x}_t, x_\diamond) \\ \Rightarrow \quad & u_t(g, \ddot{x}_t) && \leq u_t(g, x_\diamond). \end{aligned}$$

Then $x_\diamond \in \nu_t$. This concludes the proof. □

C.4 Proof for Theorem 3.6

We introduce the following lemma to help proving the theorem.

Lemma C.3. *Consider a subset of T' agents who have the same base covariate value, i.e., $\ddot{x}_t = \ddot{x}$ for all $t \in [T']$, for some value $\ddot{x} \in \mathcal{X}$. We assume that agents respond to ARexes according to Equation (3) and that Assumption 3.5 holds for all explanation methods. Then, all T' agents have the same set of non-harmful responses, i.e., $\nu_t = \nu$ for all $t \in [T']$. Moreover, for any arbitrary explanation method $(\mathcal{E}, \sigma)$ that induce a non-harmful response $x_\bullet$ from all of these agents, there exists an ARex $(\vec{x}, \hat{\vec{y}})$ that induces the same response $x_\bullet$ from these agents.*

Proof. Because these agents' responses are always identical for any explanation (Assumption 3.5), they must have the same response for any AR-based explanation. Given that AR-based explanations always induce non-harmful responses (Remark 3.3), these agents will have the same set of non-harmful responses, i.e., $\nu_t = \nu$ for all $t \in [T']$. This proves the first result.

Furthermore, for any arbitrary explanation method that induces a best response $x_\bullet \in \nu$ of these agents, it must hold that $u_t(g, x_\bullet) \geq u_t(g, \ddot{x})$ for all $t \in [T']$, because of Definition 3.1.

Consequently, there exists an AR-based explanation $(\vec{x}, \hat{\vec{y}}) = (x_\bullet, g(x_\bullet))$ and by Equation (3), all agents will follow the recommendation.

This concludes the proof. □

We now prove the theorem.

Proof for Theorem 3.6. Fix an ARex space $\mathcal{E}' = \mathcal{X} \times \mathcal{Y}$ as defined in Section 3.2 and let us define $\sigma'_{\ddot{x}} : g \mapsto \vec{x}$ as an action-recommendation function, parameterised by $\ddot{x}$. It can be seen that, for an agent population with the corresponding set of base covariates $\ddot{\mathcal{X}}_T = \{\ddot{x}_t\}_{t \in [T]}$, any ARex policy $\sigma' : (\ddot{x}, g) \mapsto (\vec{x}, \hat{\vec{y}})$ can be constructed with $\{\{\sigma'^{r}_{\ddot{x}}\}_{\ddot{x} \in \ddot{\mathcal{X}}_T}, g\}$.

Suppose that we have a specific predictive model g and an arbitrary explanation method $(\mathcal{E}, \sigma)$ that induces non-harmful responses for all T agents in the population.

Consider any subset of T' agents who have the same base covariates, i.e., $\ddot{x}_t = \ddot{x}$ for all $t \in [T']$, for some value $\ddot{x}$. Suppose that the said explanation method $(\mathcal{E}, \sigma)$ induces the non-harmful response $x_\bullet$ for these agents. Then,

following Lemma C.3, there exists an ARex $(\vec{x}, \hat{\vec{y}})$ that induces the same response $x_\bullet$. Consequently, we can construct the corresponding action-recommendation function as $\sigma'_{\ddot{x}} := \vec{x}$.

Putting together all those action-recommendation functions, we can construct an ARex policy that induce the same response tuple from the whole agent population as that induced by the said arbitrary explanation method.

This concludes the proof. □

D ADDITIONAL DETAILS OF THE EXPERIMENTS

We provide here additional details to the setups for experiments in Section 4.

D.1 On the no-harm guarantee of ARexes

We use a quartic function as the predictive model of the DM where $g(x) = x^4 - x^2 + 1$ and use 2nd-order Taylor expansions and LIME as baseline explanation methods. We generate a simple data set of 100 agents with 1-dimensional features as follows:

$$\begin{aligned} \ddot{X}_t &\sim \mathcal{N}(0, 0.4^2), \\ \alpha_t &\sim \mathcal{U}([1, 1.2]), \end{aligned}$$

where α_t denotes the cost factor of agent t, which we use to model the heterogeneity of agents' cost functions. The cost function for agent t is $c_t(\ddot{x}_t, x) = |\alpha_t| \|\ddot{x}_t - x\|_2^2$. For simplicity, we generate the AR-based explanations randomly as follows:

$$\begin{aligned} \vec{X}_t &\sim \mathcal{N}(0, 0.4), \\ \hat{\vec{Y}}_t &:= g(\vec{X}_t). \end{aligned}$$

Computational details. This experiment took less than 5 seconds to run on a standard MacBook Pro with an M2 chip and 16GB of RAM.

D.2 Operationalising ARexes on synthetic data

Synthetic data generation. We construct a synthetic dataset containing agents of 3-dimensional (observable) feature vector $\ddot{x}_t \in \mathbb{R}^3$ and a scalar (unobservable) feature $z_t \in \mathbb{R}$ as follows:

$$\begin{aligned} Z_t &\sim \mathcal{U}(\{0, 1, 2, 3\}), \\ \alpha_t \mid z_t &\sim \mathcal{N}(0.02 + 0.1 z_t,\ 0.01^2), \\ \ddot{X}_t \mid z_t &\sim \mathcal{N}(\mathbf{m},\ I \times 2), \end{aligned}$$

where $\mathbf{m} := [10 + z_t, 10 + z_t, 10 + z_t]^\top$ and the cost function $c_t(\ddot{x}_t, x) = |\alpha_t| \|\ddot{x}_t - x\|_2^2$.

Learning agents' responses. As we mention in Section 4.2, training σ^r requires the DM's ability to simulate agents' responses. Let $\hat{\psi} : (\ddot{x}_t, \hat{\ddot{y}}_t, \vec{x}_t, \hat{\vec{y}}_t) \mapsto \hat{x}_t$ denote a model that the DM can use to predict agents' responses. We construct $\hat{\psi}$ by using an underlying model $\xi : \left(\ddot{x}_t, \vec{x}_t, \Delta\vec{g}_t\right) \mapsto \hat{w}_t$ that predicts an agent's compliance w_t. We define compliance w_t as a binary variable where $w_t = 1$ indicates the agent follows the recommendation $\vec{x}_t$, and $w_t = 0$ otherwise. The term $\Delta\vec{g}_t := g(\ddot{x}_t) - g(\vec{x}_t)$ denotes the gain in prediction value for the agent t and serves as a useful feature for this classifier, given the additive structure of utility in Equation (1). Once ξ is learned, the DM can simulate an agent's response as $\hat{x}_t := \hat{w}_t \vec{x}_t + (1 - \hat{w}_t)\ddot{x}_t$.

To generate necessary data to learn the classifier ξ, the DM can employ a sampler π to generate random ARexes as follows:

$$\begin{aligned} \vec{X}_t \mid \ddot{x}_t &\sim \pi(\vec{X}_t, \ddot{x}_t), \\ \hat{\vec{Y}}_t &:= g(\vec{X}_t). \end{aligned}$$

Repeated risk minimisation. Putting everything together, with finite samples, the empirical objective in the i-th iteration of the RRM procedure is $(g_i, \sigma_i^r) = \arg\min_{g,\sigma^r}(1/T_i)\sum_{t\in[T_i]}(g(\hat{x}_t) - y_t)^2$. The notation $\hat{x}_t$ refers to the simulated agent's response based on the recommended action $\vec{x}_t := \sigma_i^r(\ddot{x}_t)$ and the inferred reaction model $\hat{\psi}$. In addition, $\{y_t\}_{t\in[T_i]}$ are the outcomes of T_i agents collected from when the previous models $(g_{i-1}, \sigma_{i-1}^r)$ are deployed, as usually done in RRM (Perdomo et al., 2020).

RRM with baseline CEs . Because the CE policy is fixed and RRM is already used, there is no need to simulate agents' responses here. Thus, the objective of this approach in each iteration of RRM is $g_i = \min_g \sum_{t\in[T_i]}(g(x_t) - y_t)^2$ where the dataset $\{x_t, y_t\}_{t\in[T_i]}$ is collected from when the previous model g_{i-1} is deployed.

The Joint-Opt algorithm. Algorithm 1 summarises the details of our Joint-Opt procedure and we explain the steps here. We use 3-layer ReLU network for constructing all three models g, σ^r, and ξ. To make the learning more efficient, we first pre-train the predictive model g and the AR function σ^r to obtain g_0 and σ_0^r. We use a dataset $D_1 = \{x_t, y_t\}_{t\in[5000]}$ for this step.

Then, we interact with the next 10^6 agents to collect another data set $D_2 = \{\ddot{x}_t, \vec{x}_t, w_t, \Delta\vec{g}_t\}_{t\in[10^6]}$. The collected dataset will later be used to train the compliance predictor ξ. To do this, we construct a sampler π to generate random ARexes:

$$\vec{X}_t \mid \ddot{x}_t \sim \mathcal{N}(x_t^\diamond, 4),$$
$$\hat{\vec{Y}}_t := g_0(\vec{X}_t),$$

where $x_t^\diamond$ is chosen arbitrarily between the following options:

$$x_t^\diamond := \ddot{x}_t, \quad \text{or}$$
$$x_t^\diamond := \sigma_0^r(\ddot{x}_t), \quad \text{or}$$
$$x_t^\diamond := \arg\min_{x\in\mathcal{X}} \left(g_0(x) + \left\|x - \ddot{x}_t\right\|_2^2\right).$$

Next, we run RRM over $m = 100$ iterations, in each of which we deploy g_{i-1} and σ_{i-1}^r to interact with 10^4 agents to collect a data set $\{x_t, y_t\}_{t\in[10^4]}$, where the subscript i denotes an iteration, as outlined in Algorithm 1. The models g_i and σ_i^r are sequentially updated over 100 iterations as specified in Equation (12) to eventually obtain the optimal g and σ^r.

Evaluation. We then compare the prediction errors between all approaches on a hold-out test set of 10^6 strategic agents. For ease of presentation, we scale the mean-squared errors by dividing them with the constant $nc = \frac{1}{T}\sum_{t\in[T]} \ddot{y}_t$ that is independent of the DM's choice of models. We refer to the scaled errors as normalised mean-squared errors (nMSE). If the loss is computed on offline data, i.e., without agents' strategic responses, we simply refer to it as nMSE, otherwise, *strategic* nMSE.

Hyperparameter choices. We report all key details necessary to understand the experimental results. Other lower-level settings (e.g., optimiser, learning rates, numbers of iterations, model sizes) are omitted from discussion as they do not affect our main conclusions. Our goal is to demonstrate the benefit of optimising the ARex policy, rather than relying on fixed designs such as counterfactual explanations. Performing extensive hyperparameter tuning for the baselines would effectively optimise those fixed policies, which would only reinforce our central claim. However, we include complete source code with the supplementary material for reproducing all experimental results.

Computational details. This experiment was completed in approximately 11 minutes on a cloud machine with 44 vCPUs and 88GB of RAM.

D.3 ARexes on German credit dataset

The dataset is publicly available at https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

Algorithm 1 Joint optimisation of g and σ.

Require: Dataset $D_1 = \{x_t, y_t\}_{t\in[T]}$ and the sampler π.
Parameters: T, T', and $\{T_1, \ldots, T_m\}$.

1: Pre-train g and σ^r on $D_1 = \{x_t, y_t\}_{t\in[T]}$ as follows:

$$g_0 = \arg\min_{g} \sum_{x_t, y_t \in D_1} (g(x_t) - y_t)^2,$$
$$\sigma^r_0 = \arg\min_{\sigma^r} \sum_{x_t, y_t \in D_1} (\sigma^r(x_t) - x_t)^2.$$

2: Interact with agents over T' rounds, with g_0 and a sampler π to collect the dataset $D_2 = \{\ddot{x}_t, \vec{x}_t, \Delta\vec{g}_t, w_t\}_{t\in[T']}$, then train the compliance predictor with the objective

$$\arg\min_{\xi} \sum_{t\in[T]} \Big(- w_t \log(\hat{w}_t) - (1 - w_t) \log(1 - \hat{w}_t) \Big),$$

where $\hat{w}_t$ is the output of ξ and w_t is the actual label.

3: **for** $i \in \{1, \ldots, m\}$ **do**

4: Interact with agents over T_i rounds with g_{i-1} and σ^r_{i-1} to collect the dataset $D_{3,i} = \{\ddot{x}_t, y_t\}_{t\in[T_i]}$

5: Update (g_i, σ^r_i) by solving

$$\big(g_i, \sigma^r_i\big) := \arg\min_{g, \sigma^r} \sum_{t\in[T_i]} \big(g(\hat{x}_t) - y_t\big)^2, \tag{12}$$

where $\hat{x}_t$ is simulated with the compliance predictor ξ and the input $\ddot{x}_t$.

6: **end for**

7: Set $(g, \sigma^r) := (g_m, \sigma^r_m)$.

As mentioned in Section 4.3, we adopt details from Xie and Zhang (2024) for pre-processing the data. Specifically, we remove two sensitive features (i.e., *age* and *sex*) and designate 8 out of the remaining 18 features as modifiable for strategic agents, these are: *existing account status, credit history, credit amount, savings account, present employment, installment rate, guarantors*, and *residence*. Categorical features are label encoded, for simplicity, and numerical features are standardised to have zero mean and unit variance. For each modifiable feature (indexed with i), we identify the range of feasible values and denote it as $[x_i^L, x_i^U]$. We use this to prevent the DM from recommending extreme feature changes to agents and to prevent agents from adopting such extreme feature modifications.

As mentioned in Section 4.3, we design the cost function as

$$c(\ddot{x}_t, x) := 0.01 \sum_{i\in\mathcal{I}} \frac{|\ddot{x}_{ti} - x_i|}{(x_i^U - x_i^L)},$$

where $\mathcal{I}$ is the index set of modifiable features. Any change in a non-modifiable feature incurs infinite cost. We use the weighted $L1$ distance instead of the quadratic form to avoid the cost values from becoming excessively small. Furthermore, the scaling factor of 0.01 is heuristically chosen so that we have a nice balance between agents who can easily change their features and agents who cannot. This choice enables us to more clearly observe the impact of different designs for ARex policies.

In this dataset, the agent's outcome y_t is a binary variable indicating a customer's credit risk classification (i.e., 1 means *good* and 0 means *bad*). To simulate how an agent's outcome y_t changes w.r.t. their strategically updated covariate x_t, we use a logistic regression model that is fitted on this dataset of 1,000 observations. Let $s : \mathcal{X} \to (0, 1)$ be the resulting logistic function that outputs the probability of the outcome y_t is positive. We simulate the agent's outcome as

$$Y_t \mid x_t \sim \text{Bernoulli}\big(s(x_t)\big).$$

We fit CTGAN (Xu et al., 2019) on the original dataset then generate additional samples: 9,000 for training and 1,000 for testing.

Evaluation. Counterfactual explanations are generated similarly to the synthetic setup, with additional constraints (e.g., categorical/bounded features) enforced via projected gradient descent. Unlike previous setup, we now use $-g_s(x)$ in the objective function for generating CEs as the agents now benefit from higher prediction scores. The training of g follows the same procedure as before, except for the loss functions: we use *weighted* binary cross-entropy (BCE) to learn agents' response function ξ (with imbalanced labels), and *unweighted* BCE in RRM, where outcomes y_t may shift due to strategic behavior. We evaluate predictive accuracy (using the F_1 score) on the hold-out test set of 1,000 strategic agents, referring to this as *strategic* F_1 *score* to reflect possible shifts in (x_t, y_t).

Hyperparameter choices. Similar to the previous synthetic setup, we report all key details necessary to understand the experimental results. Other lower-level settings (e.g., optimiser, learning rates, numbers of iterations, model sizes) are omitted from discussion as they do not affect our main conclusions. We include complete source code with the supplementary material for reproducing all experimental results.

Computational details. This experiment was completed in approximately 4 minutes on a cloud machine with 44 vCPUs and 88GB of RAM.

D.4 ARexes on the RSNA Pediatric Bone Age dataset

We use the 2017 RSNA Pediatric Bone Age dataset (Halabi et al., 2018) that contains x-ray images of infants' hands and the task is to predict their bone age (in months). The training set consists of $12,611$ images and the test set contains $1,425$ images. The dataset is publicly available at https://www.rsna.org/rsnai/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017.

Scenario. We consider the scenario in which a hospital (acting as a DM) wants to acquire more x-ray data to help improve their predictive model g, and clinics (acting as strategic agents) want to sell their data.

A clinic t submits an x-ray image $\ddot{x}_t$ alongside with the recorded patient's outcome $\ddot{y}_t$. The hospital has the ability to evaluate the quality of this data and to pay the clinic an amount specified by $b(g, \ddot{x}_t, \ddot{y}_t)$. Let $\kappa : \mathcal{X} \to \mathbb{R}^k$ denotes a function that associate an x-ray image x with a vector of k their properties. For example, this could contain the x-ray machine's settings when the image was taken such as *exposure*, *focus*, or *contrast*. After evaluating the submitted x-ray image $\ddot{x}_t$, the hospital could give a recommendation to the clinic on how to obtain higher payment, i.e., b', by re-submitting another image. This recommendation can come from the form of the recommended properties $\kappa' \in \mathbb{R}^k$. The clinic can look for another image x_t such that $\kappa(x_t) = \kappa'$.

After submitting the image x_t (and the associated patient outcome y_t), the clinic receives the payment $b(g, x_t, y_t)$. The objective of the hospital is to minimise the expected predictive error of the model g, while taking into account the strategic behaviour of the clinics:

$$\min_{g,\sigma} \mathbb{E}\left[(g(X_t) - Y_t)^2\right],$$

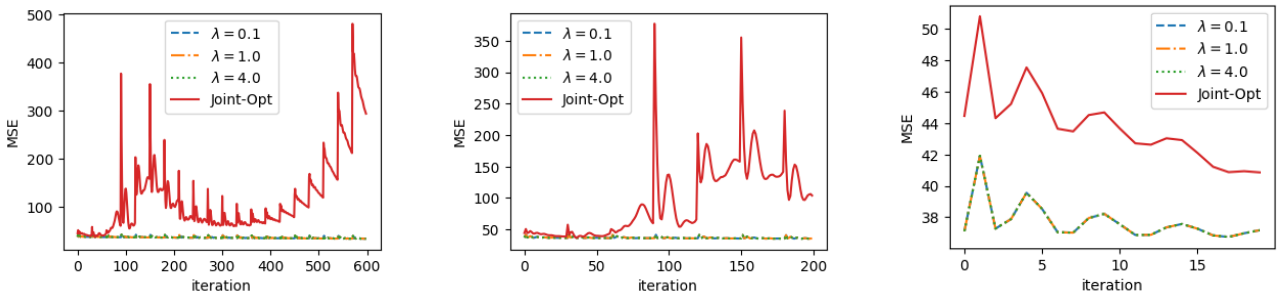
where $\sigma : (g, \ddot{x}) \mapsto (\kappa', b')$ denotes the ARex policy.

Setup details. Following Lu et al. (2024), we create the embeddings for the images using a pretrained CLIP ViT-B/32 model. We denote the embedding of an image x as $\phi(x) \in \mathbb{R}^{512}$. To simulate the change in outcome y when the covariate x changes, we assume a linear relationship between these two, in the embedding space, and fit a linear model h, where:

$$y := h(x) := \phi(x)^\top \beta + \beta_0.$$

Similar to the previous two experiments, we use 3-layer ReLU networks to construct the models g, σ^r, ξ. For simplicity, we assume here that the properties of an image $\kappa(x)$ coincide with the first k entries in their embedding vector $\phi(x)$. Then, we can define the ARex policy σ and action-recommendation function σ^r as follows:

$$\sigma : (\ddot{x}, g) \mapsto (\kappa', b'),$$
$$\sigma^r : \ddot{x} \mapsto \kappa'.$$



(a) From 0th to 600th iteration. (b) From 0th to 200th iteration. (c) From 0th to 20th iterations.

Figure 3: The plots show the training errors of our Joint-Opt approach and the baselines at different magnification levels. The training error of Joint-Opt increases **more** sharply after each interaction between the DM and agents. This implies an unstable optimisation process and the absence of fixed-point convergence (Perdomo et al., 2020), especially in high-dimensional settings where the number of trainable parameters increases significantly due to joint-training of $\{g, \sigma^r\}$.

We set the payment function as $b(g, x, y) := (g(x) - y)^2$ to reward an agent an amount that corresponds to how much the new data point surprises the current predictive model g. For simplicity, we assume that agents operate directly on the embedding space and their cost functions have the form of $c_t(x, \ddot{x}_t) := \alpha_t \|\phi(x) - \phi(\ddot{x}_t)\|_2^2$, where α_t represents the heterogeneous factor of this cost function.

Baselines and evaluation. Similar to the previous two experiments, we use fixed designs of counterfactual explanations as baselines and evaluate the performance of g with mean squared errors (MSE).

Computational details. This experiment was completed in approximately 5 minutes on a machine with an NVIDIA A100 GPU, 12 vCPUs and 85GB of RAM.

Results. Figure 3 shows the training errors of our Joint-Opt approach and the baselines at different magnification levels. The training error of Joint-Opt increases **more** sharply after each interaction between the DM and agents. This implies an unstable optimisation process and the absence of fixed-point convergence (Perdomo et al., 2020), especially in high-dimensional settings where the number of trainable parameters increases significantly due to joint-training of $\{g, \sigma^r\}$.

This phenomenon reflects the coupled nature of joint training: as the ARex policy σ is updated, it shifts the distribution of agents' covariates P_X, which in turn affects the optimisation of g. When the parameter space is large, these feedback effects make convergence difficult. The issue is further compounded by the choice of the DM's payment function b and the agents' cost functions c_t, both of which strongly influence stability. As pointed out by Perdomo et al. (2020), RRM can only converge under some conditions on the loss function.

Consequently, while our joint-optimisation procedure performs reliably in low- to medium-dimensional settings, extending it to high-dimensional and more complex environments remains an open and interesting direction. A systematic study of convergence behaviour in such regimes would complement our current focus. Related work by Tsirtsis and Gomez Rodriguez (2020) also explores joint-optimisation but restricted to discrete and classification settings.

Overall, our focus lies in establishing a theoretical foundation **to characterise explanations** in terms of their impact of agents' behaviour. While we prove the existence of ARexes and ARex methods for sufficiently inducing no-harm actions in agents, it would be an interesting future direction to thoroughly investigate how one can optimally design an ARex policy in a high-dimensional setting.

E RELATED WORK (EXTENDED)

Strategic learning. Strategic classification was introduced by Brückner et al. (2012) and further developed by Hardt et al. (2016), where they presented the first computationally efficient algorithms to learn near-optimal classifiers in strategic environments. Their key assumption was that agents have complete knowledge of the classifier due to information leakage, even when the system is designed to obscure the model. In contrast, our work weaken this assumption by considering scenarios where the learner (or DM) releases partial information

about their model.

Subsequent research has expanded the field of strategic classification by developing more efficient algorithms (Dong et al., 2018; Levanon and Rosenfeld, 2021; Ahmadi et al., 2021) or by incorporating new aspects such as social welfare (Hu et al., 2018; Milli et al., 2019), randomisation (Sundaram et al., 2023; Ahmadi et al., 2023; Shao et al., 2024), repeated interactions (Harris et al., 2021; Cohen et al., 2023), and incentivising improvements (Kleinberg and Raghavan, 2020; Harris et al., 2022b; Vo et al., 2024). Another important thread considers settings where agents cannot best respond to the true model—either due to bounded rationality, limited information, or uncertainty in their response process (e.g., (Jagadeesan et al., 2021; Bechavod et al., 2022; Harris et al., 2022a; Xie et al., 2024)). Our work is complementary: instead of committing to a particular agent model, we focus on how the DM can structure the information disclosed through explanations, and we conduct a comparative analysis across explanation types to understand when they induce responses that do not harm agents.

In particular, the works of Harris et al. (2022a) and Cohen et al. (2024) are closest to ours. As discussed throughout Section 3.2, Harris et al. (2022a) focuses on the obedience-inducing property (also known as the Bayesian incentive compatibility) of a subclass of action recommendations, whereas we focus on the no-harm property of action recommendations. As we also discuss in Section 3.2, identifying an AR-based explanation policy that can induce obedience for each *individual* agent is hard, especially when agents are heterogeneous (e.g., in Harris et al. (2022b); Shao et al. (2024)), and such *individual*-level identification might not be necessary if the DM only cares about optimising their expected utility, which is computed over the population of agents. Unlike action recommendations, Cohen et al. (2024) instead releases a subset of the hypothesis class to all agents, aligning with global explanations in our framework (Section 2). However, interpreting a set of models—such as neural networks—can be difficult for agents. In contrast, the AR-based explanations are not only more interpretable but also provide guidance that cannot mislead agents.

Counterfactual explanations and algorithmic recourse. Tsirtsis and Gomez Rodriguez (2020); Karimi et al. (2022) explore counterfactual explanations and algorithmic recourse, for strategic agents. Although algorithmic recourse focuses on recommending actions to achieve better outcomes, it actual implementation often requires strong causal assumptions. These assumptions can render it impractical in more general settings where such causal knowledge is not justified. In contrast, our work adopts a weaker notion of desirability centred on agents' welfare—ensuring non-harmful responses—and examines a broader range of explanation types beyond counterfactuals. Even though both Tsirtsis and Gomez Rodriguez (2020) and our work involve counterfactual explanations, the contributions differ. Precisely, they focus on optimising CEs in strategic settings, while we analyse multiple explanation types and formally show why ARexes are more desirable. In addition, our proposed learning procedure in Section 4.2, though not the main focus, is designed to be more general, extending beyond the classification and discrete case in Tsirtsis and Gomez Rodriguez (2020).

Information design. The extensive literature on information design, as surveyed by Bergemann and Morris (2019), studies how to design information disclosure policies in a game of two parties. While our results are inspired by these works, e.g., Theorem 3.6, the goals differ significantly. As discussed in Section 3.2, information design aims at *persuading* agents with a general response model and does not necessarily ensure the no-harm property (Definition 3.1). In contrast, we study explanation methods that prioritise the no-harm property, ensuring agents' welfare is not compromised. By incorporating specific agent models in strategic settings, we establish the sufficiency of AR-based explanations without requiring the DM to account for agents' heterogeneous reaction models.