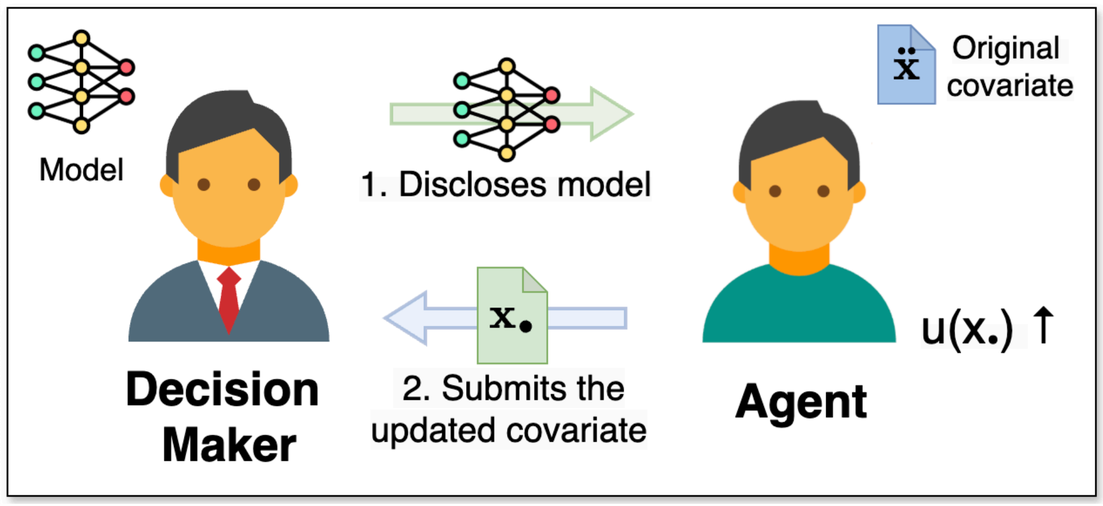
EXPLANATION DESIGN IN STRATEGIC LEARNING:
Sufficient Explanations that Induce Non-harmful Responses

Kiet Q.H. Vo[1], Siu Lun Chau[2], Masahiro Kato[3], Yixin Wang[4], Krikamol Muandet[1]

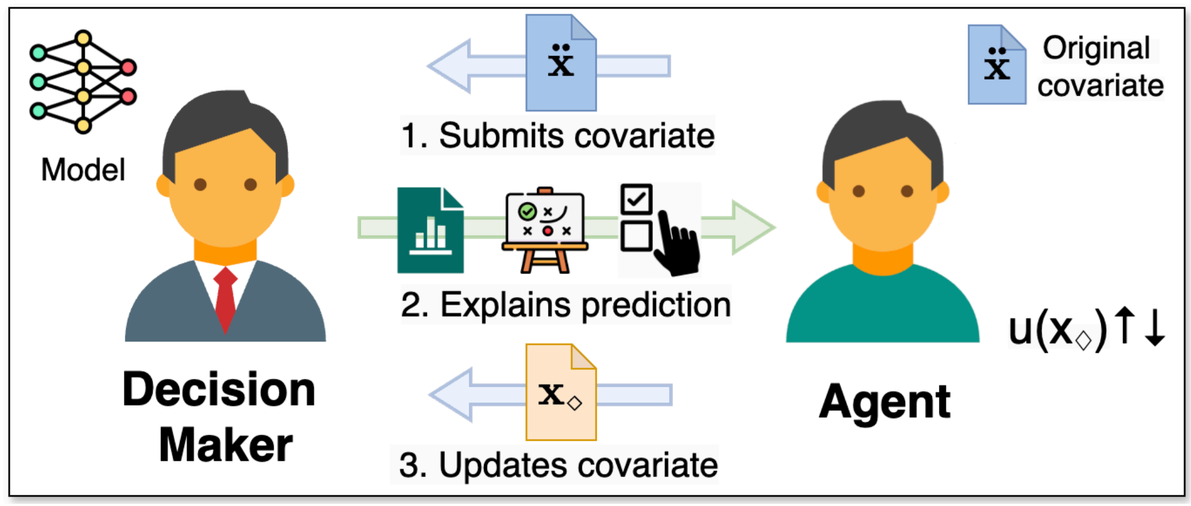
AN EXPLANATION DESIGN PROBLEM

Full Disclosure of Predictive Model



Here, the agent can *correctly anticipate* how changing $\ddot{x}$ affects the prediction, then picks an update $x_\bullet$ that *reliably improves* utility $u(x_\bullet)$.

Partial Disclosure via Explanations (Ours)



With only an explanation (i.e., partial information), the agent's update $x_\diamond$ *might not improve* utility $u(x_\diamond)$.

Q1: Can we ensure no reduction in agents' utilities?
Q2: Is there a sufficient class of explanations that guarantee this?

YES!

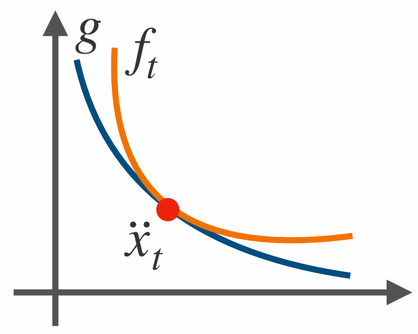
SETUP

0. **Agent** is realised: $(\ddot{x}_t, c_t) \sim P_{\ddot{X},C}$, where c_t is a cost function.
1. **DM** predicts agent's risk: $g(\ddot{x}_t)$.
2. **DM** gives explanation: $e_t := \sigma(g, \ddot{x}_t)$.
3. **Agent** modifies covariate: $x_t := \psi(\ddot{x}_t, e_t, c_t)$.
4. **DM** updates the prediction from $g(\ddot{x}_t)$ to $g(x_t)$.

Agent's true utility: $u_t(g, x) := b(x) - c_t(\ddot{x}_t, x)$
$= -g(x) - c_t(\ddot{x}_t, x)$.

Agent's non-harmful responses:
$\nu_t = \{x \in \mathcal{X} : u_t(g, x) \geq u_t(g, \ddot{x}_t)\}$.

KEY CONTRIBUTIONS



A *necessary condition* to ensure surrogate models do not mislead agents into self-harming actions.

Action recommendation (AR)-based explanations (ARexes) make up *a sufficient class* that guarantees non-harmful agents' responses.

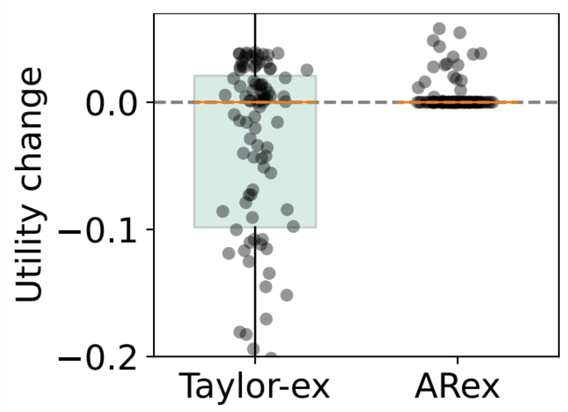
SURROGATE MODELS

- If it holds that, every cost function c_t induces a response $x_t \in \nu_t$, then f_t *must satisfy*:
  $f_t(\ddot{x}_t) - f_t(x') \leq g(\ddot{x}_t) - g(x')$
  $\forall x' \in \mathcal{X}_t^{g\downarrow}$.
- Corollary 3.3 extends this result to other forms of explanations.
- The surrogate model *must understate* the agent's gain in prediction score.

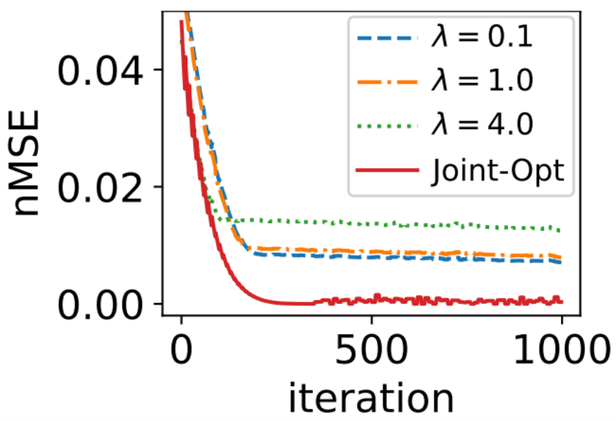
AR-BASED EXPLANATIONS

- For any *arbitrary* explanation e' that induce $x_\bullet \in \nu_t$,
  There exists an ARex $(\vec{x}_t, \hat{\vec{y}}_t)$ that induces the *same response* $x_\bullet \in \nu_t$.
- Theorem 3.6 extends this result to ARex methods for heterogeneous agents.
- It suffices to *use only* AR-based explanations.

EXPERIMENTS



ARexes do not reduce agents' utilities.



In practice, ARexes can be jointly optimised with predictive model g.

CISPA HELMHOLTZ-ZENTRUM FÜR INFORMATIONSSICHERHEIT
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
MIZUHO
UNIVERSITY OF MICHIGAN