Designing Algorithmic Delegates

The Role of Indistinguishability in Human-AI Hand-Off

Sophie Greenwood[1], Karen Levy[1], Solon Barocas[1,2], Hoda Heidari[3], and Jon Kleinberg[1]

[1] Cornell University [2]Microsoft Research [3]Carnegie Mellon University

Introduction

Humans are increasingly willing to **delegate** decisions to AI agents

A human decides whether to delegate based on properties of the specific **instance** of the decision-making problem they face
Humans lack full awareness of all the factors relevant to this choice

- They perform a kind of **categorization** by treating **indistinguishable** decision-making instances as the same
- [Example: a human using an online shopping bot may know that an item is rare but not whether prices are higher than usual. When deciding whether to delegate, the human must group these different possibilities into one "rare" category.]

We examine the **tractability** of **designing the optimal algorithmic delegate** in the presence of categorization

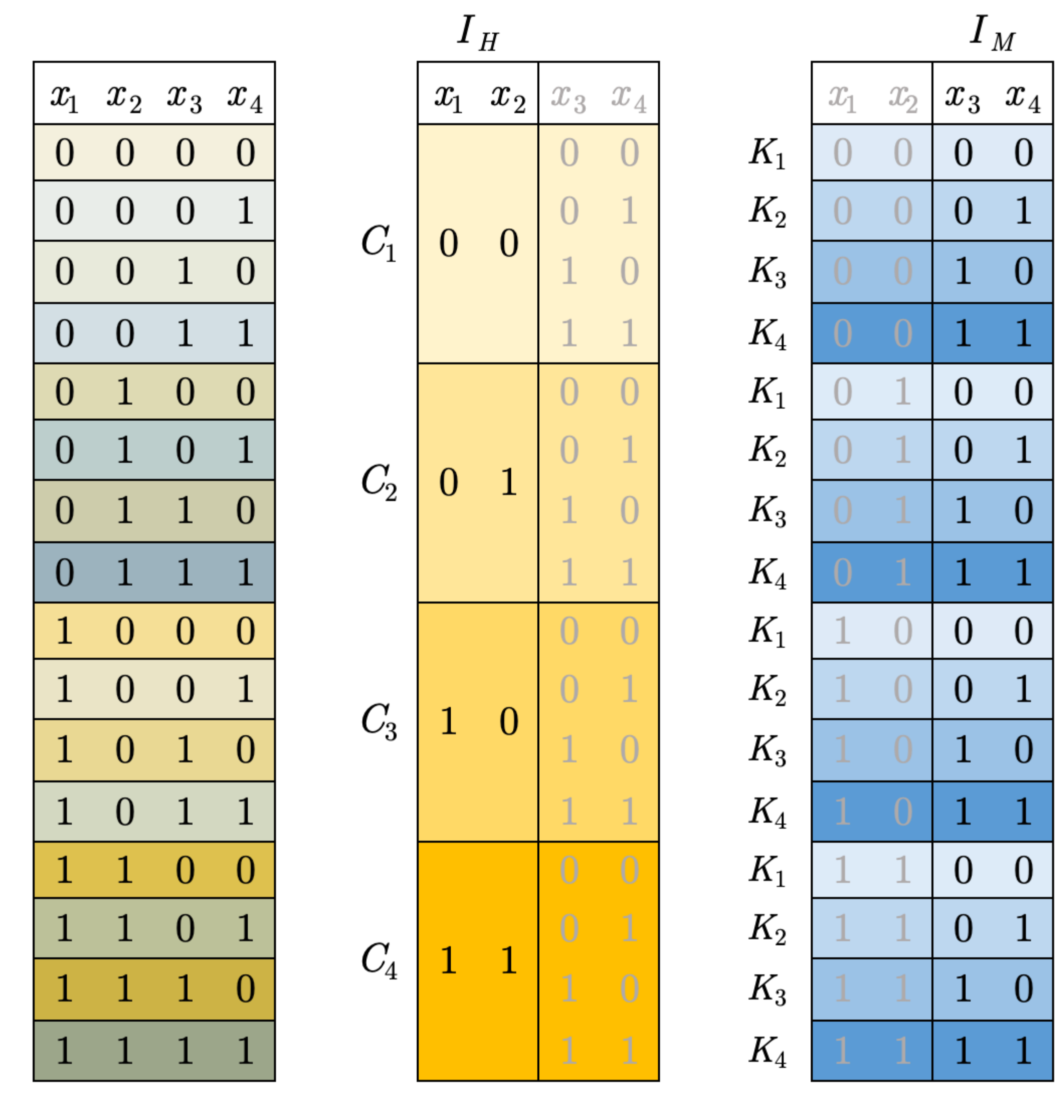

Figure 1. Visualization of states and categories

Model

Assume (for now): each state occurs with equal probability

Assume (for now): I_H and I_M partition $[d]$

d binary features $x_1, x_2, \dots, x_d \in \{0,1\}$; $n = 2^d$ states of the world $\mathbf{x} = (x_1, x_2, \dots, x_d)$
In each state $\mathbf{x}$ there is some ground truth optimal action $f^*(\mathbf{x})$
Decision y in state $\mathbf{x}$ has loss $\left(y - f^*(\mathbf{x})\right)^2$

Two agents: a **human** and a **machine**. The human observes features I_H; the machine observes features I_M
A **human category** C is a set of states that are indistinguishable to the **human**
A **machine category** K is a set of states that are indistinguishable to the **machine**

Faced with category C, the human selects $f_H^*(C) \coloneqq \mathbb{E}[f^* | C]$ $\longrightarrow$ loss $\ell_H(f_H^*, C) \coloneqq \mathbb{E}\left[\left(f_H^*(C) - f^*(\mathbf{x})\right)^2 \middle| \mathbf{x} \in C\right]$

Faced with category K, the machine selects $f_M(K)$ $\longrightarrow$ loss $\ell_M(f_M, C) \coloneqq \mathbb{E}\left[\left(f_M(K(\mathbf{x})) - f^*(\mathbf{x})\right)^2 \middle| \mathbf{x} \in C\right]$

Delegation process:

Human observes $C(\mathbf{x})$, and $\ell_M(f_M, C) \gtrless \ell_H(f_H^*, C)$

$>$: delegate → Machine observes $K(\mathbf{x})$, and selects $f_M(K)$

$\leq$: do not delegate → Human selects $f_H^*(C)$

$f_M(K)$: the machine designer's choice

Team loss: $\ell(f_H^*, f_M) \coloneqq \sum_C \min\{\ell_H(f_H^*, C), \ell_M(f_M, C)\}$

Options for f_M:

Oblivious machine: $f_M^{\text{obliv}}(K) \coloneqq \mathbb{E}[f^* | K]$

Optimal machine: $f_M^*(K) \coloneqq \min_{f_M} \ell(f_H^*, f_M)$

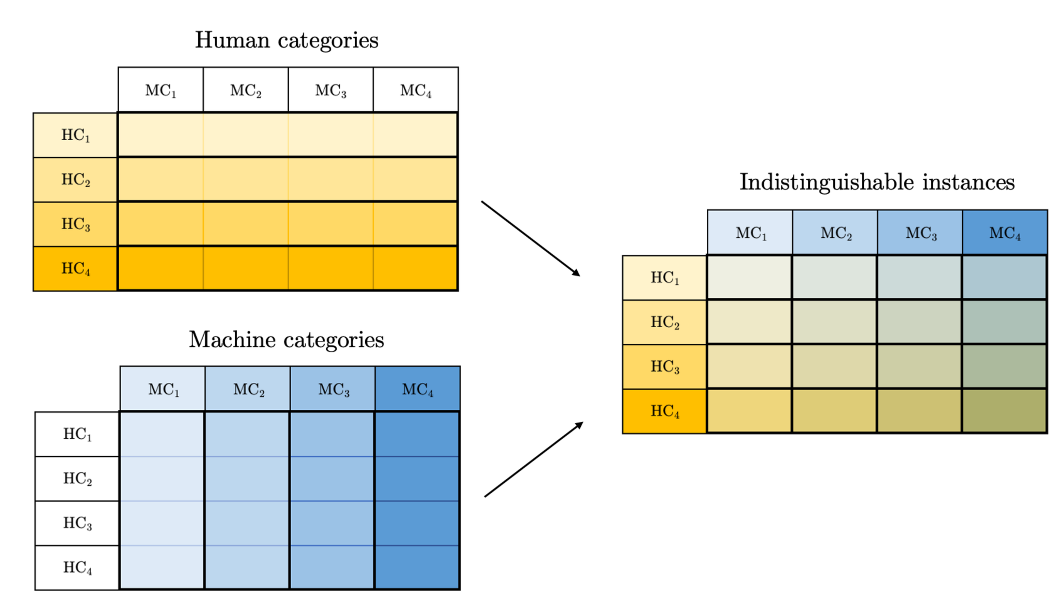

Figure 2. Visualization of categories in a grid

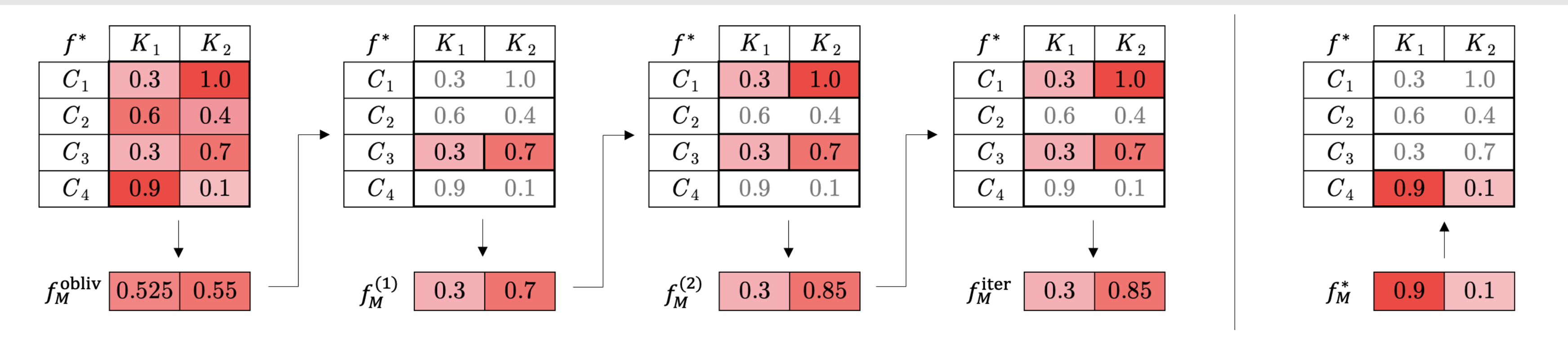

Figure 4. Illustration of iterative design in an example setting. Iteratively reoptimize the oblivious delegate to perform well in categories where it is adopted. This **improves** the delegate but is **suboptimal**.

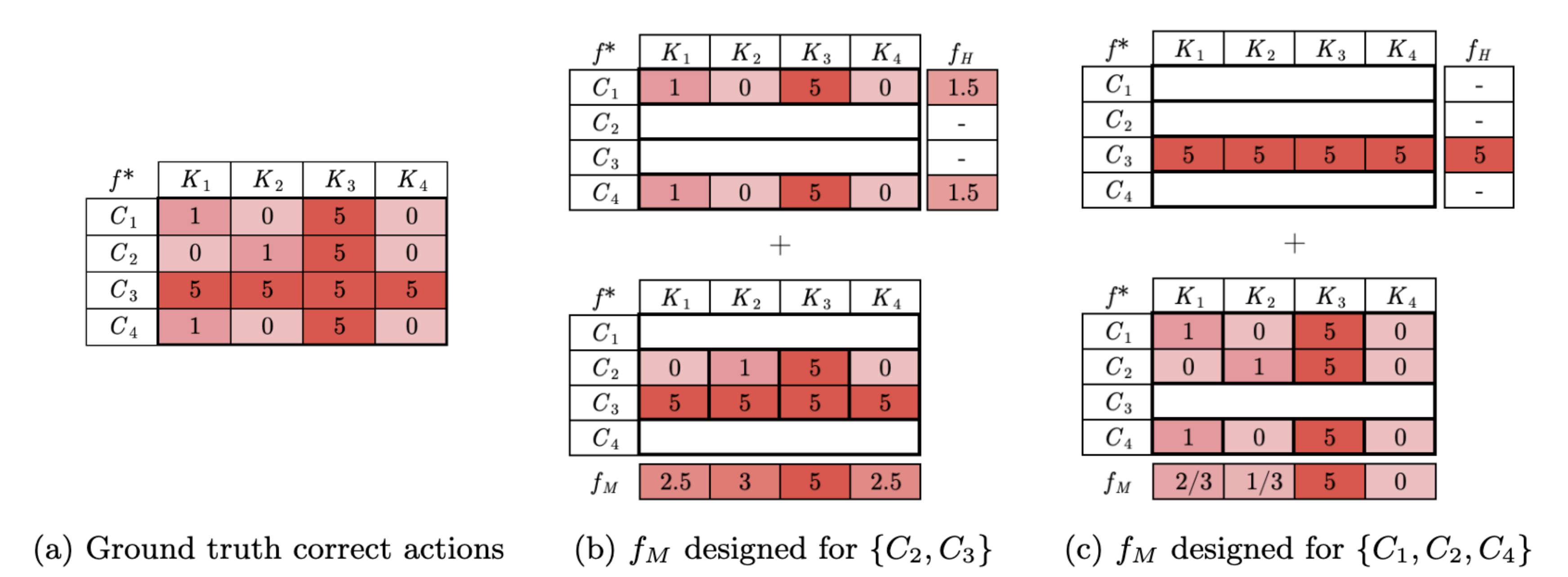

Figure 5. Illustration of optimal design; (b) shows a suboptimal solution, and (c) shows the optimal solution. The optimal machine is designed to operate in high-variance rows that have low variance across columns.

Combinatorial Reformulation

For a set of "retained" human categories $\mathcal{R}$, let $f_M^{\mathcal{R}}(K) \coloneqq \mathbb{E}[f^* | X(\mathcal{R}) \cap K]$, where $X(\mathcal{R})$ is the set of states in $\mathcal{R}$

Proposition 1 (Informal). To find an optimal delegate f_M^*, it is **sufficient and necessary** to find a set of human categories $\mathcal{R}$ that attains the minimum team loss when the human delegates to $f_M^{\mathcal{R}}$ in precisely the categories in $\mathcal{R}$.

Put simply: the problem reduces to finding the best $f_M^{\mathcal{R}}$, assuming it's adopted in $C \in \mathcal{R}$

This can be expressed as the following especially clean **combinatorial problem**:

Proposition 2. Define a matrix V with entries $v_{ij} = f^*(\mathbf{x}_{ij})$. The problem of finding an optimal delegate is as follows:

Variance Assignment. Fix a set of rows S of V. For each row $i \in S$, pay a cost proportional to the variance of V across row i, and remove row i from V. Then, for each column j, pay a cost proportional to the variance across column j of the remaining entries. Find a set S^* that minimizes the total cost.

Then for $\mathcal{R} = \{C_i : i \notin S^*\}$, $f_M^{\mathcal{R}}$ will be an optimal delegate.

Tractability

The input size of the problem is n, as we must specify $f^*(\mathbf{x})$ for each of the n states $\mathbf{x}$.

Finding the optimal delegate is tractable when the optimal action function f^* is additively decomposable into function of the human and machine categories.

Theorem 3. Suppose that f^* is **separable**, that is,
$$f^*(x) = u(C(\mathbf{x})) + w(K(\mathbf{x}))$$
for some functions u, w.
Then we can find an optimal delegate f_M^* in time polynomial in n.

In particular, **linear** functions are separable.

The problem is also tractable if the human or machine has a small number of features.

Theorem 4. Suppose that $|I_H| = O(1)$ or $|I_M| = O(1)$.
Then we can find an optimal delegate f_M^* in time polynomial in n.

However, the problem is NP-hard in general.

Theorem 5. Unless P = NP, there is **no algorithm** to find an optimal delegate f_M^* in time polynomial in n for all ground truth functions f^*.

This is because **Variance Assignment** is NP-hard.

Implications

Designing the optimal delegate is fundamentally a **hard combinatorial problem**

In **separable** settings or settings where one agent has **only a few features**, we can efficiently compute the optimal delegate

Extensions

- General **distributions** over states
- Arbitrary **categories** and **feature configurations**
- Characterizing optimal delegates in **two-feature** settings
- Computational experiments on **iterative design**

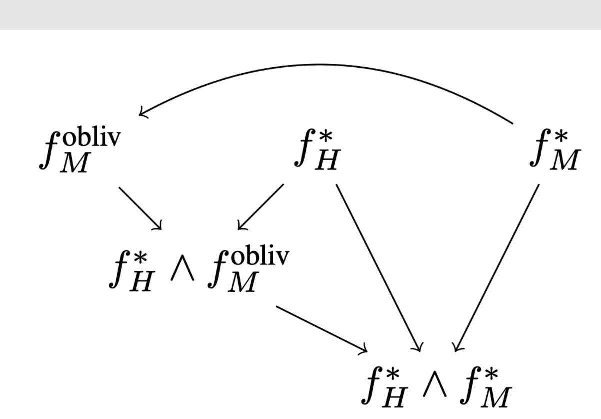

Figure 3. Relationships between the losses of different human and machine teams