arXiv:2405.01484v3 [cs.HC] 19 Nov 2026

Designing Algorithmic Recommendations to Achieve Human–AI Complementarity

Bryce McLaughlin
Wharton

Jann Spiess
Stanford GSB

First public version: January 2024
This version: November 2026

Abstract

Algorithms frequently assist, rather than replace, human decision-makers. However, the design and analysis of algorithms often focus on predicting outcomes and do not explicitly model their effect on human decisions. This discrepancy between the design and role of algorithmic assistants becomes particularly concerning in light of empirical evidence that suggests that algorithmic assistants again and again fail to improve human decisions. In this article, we formalize the design of recommendation algorithms that assist human decision-makers without making restrictive ex-ante assumptions about how recommendations affect decisions. We formulate an algorithmic-design problem that leverages the potential-outcomes framework from causal inference to model the effect of recommendations on a human decision-maker's binary treatment choice. Within this model, we make a connection between algorithmic recommendations and instrumental variables. Leveraging this connection, we introduce a monotonicity assumption that leads to an intuitive classification of human responses to the algorithm. Under this assumption, we can express the human's response to algorithmic recommendations in terms of their compliance with the algorithm and the active decision they would take if the algorithm sends no recommendation. We showcase the utility of our framework using an online experiment that simulates a hiring task. We argue that our approach can make sense of the relative performance of different recommendation algorithms in the experiment and can help design solutions that realize human–AI complementarity. Finally, we leverage our approach to derive maximin-optimal recommendation algorithms that can be implemented with machine learning using limited training data.

Bryce McLaughlin (bryce.mclaughlin@bc.edu), Carroll School of Management, Boston College and Jann Spiess (jspiess@stanford.edu), Graduate School of Business, Stanford University. We thank Maya Balakrishnan, Haosen Ge, Yagan Hazard, Sanmi Koyejo, Carl Meyer, Adrienne Propp, Emma Rockall, Omer Shiran-Cohen, Jason Weitze, and audience members at the 2023, 2024 & 2025 INFORMS Annual Meetings, 2024 MSOM Conference, 2025 Machine Learning in Economics Summer Institute, 2025 AOM Annual Meeting, Stanford, Wharton, Berkeley, MIT, Boston University, HBS, Columbia, and NYU for helpful comments and suggestions. We used AI tools (Claude Opus 5.5 and Claude Fable 5.1) in November 2026 for help with editing and the development of additional proofs. All AI edits were reviewed by the authors.

1 Introduction

Algorithms often assist, rather than replace, human decision-makers. For example, in the United States, federal pretrial service officers set bail for defendants after observing a pretrial risk assessment (Cadigan and Lowenkamp, 2011), rather than letting the algorithm take a bail decision automatically. Similarly, doctors may be assisted by clinical decision support systems throughout the patient care process (Sutton et al., 2020) and managers may be assisted by pre-employment assessments throughout the hiring process (Raghavan et al., 2020). In all of these cases, the human maintains the final decision authority. However, the design and analysis of these algorithms often focus on the prediction of outcomes as if algorithmic recommendations were implemented directly, rather than on how these recommendations affect human decisions.

The narrow focus of existing approaches on predicting outcomes rather than optimizing for human–AI complementarity becomes of particular concern in light of a growing body of evidence that shows that human decision-makers with access to algorithmic recommendations often make worse decisions than if these recommendations had been implemented directly and, in some cases, may even make worse decisions than they would have taken without access to the algorithm. Hemmer et al. (2021) identifies 53 articles that report (i) unassisted human performance, (ii) algorithm performance, and (iii) algorithm-assisted human performance on a decision task. Only in 16 of these does the algorithm-assisted human decision-maker perform best. The meta-analysis of Vaccaro, Almaatouq, and Malone (2024) finds that on average human–AI combinations perform worse than the better of human and AI alone, with the losses concentrated in decision tasks. In many of these studies, decision-makers exhibit noisy and biased adoption of the algorithm's recommendations, which counteracts the gains associated with the additional information the algorithm provides. This suggests the necessity to incorporate decision-makers' adoption decisions explicitly when designing the algorithm.

This article formalizes the design of recommendation algorithms that assist humans in making binary treatment choices via an optimization problem that leverages the potential-outcomes framework from causal inference. To analyze recommendation algorithms, we employ the notion of "potential decisions" and make connections to the analysis of instrumental variables in causal inference with imperfect compliance. This view suggests that an optimal recommendation algorithm should take into account the final outcome of interest to which the decision is related *as well as* the decision-maker's choice of when to comply with a recommendation. We then show that this framework helps design better recommendation algorithms and make sense of their empirical performance through an online experiment. Finally, we derive robust recommendation algorithms that achieve maximin performance from limited training data under varying assumptions about the compliance behavior of the decision-maker. While our theorems and propositions explicitly hold for this stylized setup, the framework we develop is broadly applicable for designing algorithms that assist human decision-makers. Specifically, we connect algorithm design choices to assumptions

adopted from the instrumental-variables literature, exposing the central role of the reaction of the human decision-maker to algorithmic recommendations.

In Section 2, we introduce a *causal design problem with endogenous response* that formalizes the objective of AI recommendation systems that assist a human decision-maker in making treatment choices. In this model, an algorithm sends a recommended action to a decision-maker. The decision-maker then chooses a binary action, which we also call the treatment, before some outcome of interest is realized. For example, the decision-maker could be a doctor that receives an algorithmic recommendation of whether to test a patient or not. Or she could be a judge who reacts to the output of a risk assessment tool. Or the decision-maker could be a hiring manager who decides whether or not to employ an applicant with the help of an algorithmic hiring recommendation. In all of these cases, we assume that the designer's goal is to maximize some beneficial outcome such as the patient's health, public safety, or the firm's productivity. We allow the decision-maker to have additional, private information that the algorithm does not see, and we also allow for complementarity in information and expertise between the algorithm and the decision-maker, thus moving beyond models in which perfect adherence to the algorithmic recommendation is optimal. Our approach builds upon a classical treatment choice or policy learning problem where the algorithm directly chooses the action rather than leaving it to the human decision-maker.

Relative to a standard policy learning problem where an algorithm directly implements treatment choices, solving our algorithmic design problem requires accounting for endogenous decisions by the human decision-makers to the recommendations. Section 3 therefore connects our problem setup to the instrumental-variables framework from causal inference, where we use the potential-outcome notation to express how decisions vary with different recommendations. Using our "potential decisions", we can state the optimal recommendation algorithm as maximizing an objective that separates into two parts, (1) the expected outcome under a direct implementation of the algorithm's recommendations and (2) the gain (or loss) due to how decision-makers deviate from the recommendations. We then introduce a monotonicity assumption that allows us to describe the effect of recommendations in terms of compliance, similar to the seminal analysis of Imbens and Angrist (1994) for instrumental variables. This monotonicity assumption simplifies the analysis of endogenous responses since it implies that recommendations can only move the decision-maker toward the recommended action. It also allows us to characterize human choices by a combination of their active decision they would take absent a recommendation as well as their compliance behavior. As a result of this simplification, we can state our optimization problem in terms of a functional of active decisions and compliance alone, further linking the design of recommendations to the choice of optimal instruments.

In Section 4, we showcase our framework by applying it to an online experiment that involves a hypothetical hiring task. This experiment asks subjects to make 25 hiring decisions and randomly varies the recommendation algorithm across subjects. This setup allows us to evaluate how the

structure of recommendation algorithms impacts ultimate task performance. Specifically, the separate recommendation algorithms are based on different assumptions about compliance behaviors. In line with our framework, we find that subjects perform better when the recommendations are designed to provide complementary information, rather than optimized for hiring decisions directly. The best-performing algorithm assumes that human decision-makers comply with the algorithm selectively, only impacting the cases where the human is uncertain. The decisions assisted by this algorithm outperform unassisted human decisions as well as the best performance by an algorithm on its own. We thus demonstrate that our framework can help design systems that improve decisions through human–AI complementarity.

Finally, in Section 5, we leverage our potential-decision framework to find robust recommendation algorithms from limited data. Here, we map different assumptions about available data and potential decisions to maximin optimal recommendations. We first focus on the realistic case where the designer may only have data about unassisted human decisions available and thus has to rely on assumptions about compliance and active decisions. Assuming that the decision-maker is more likely to comply in cases where they do not make good decisions themselves, we show that learning successful recommendations is about *learning from mistakes* rather than from the relationship of outcomes to characteristics alone. We then consider an alternative approach that is motivated by the connection to instrumental variables: an intent-to-treat (ITT) strategy based on randomized recommendations. Across the approaches, we end by discussing the implementation of the resulting robust optimization program using machine learning.

Our work responds to an empirical literature that analyzes inefficiencies in the interactions between human decision-makers and prediction algorithms. This literature has identified cases where human decision-makers imperfectly use prediction algorithms in high-stakes decisions in criminal justice (Imai et al., 2021; Stevenson and Doleac, 2022; Angelova, Dobbie, and Yang, 2023), healthcare (Agarwal et al., 2023; Maron et al., 2020), and child protective services (Fogliato, De-Arteaga, and Chouldechova, 2022; Mills and Grimon, 2022), so much so that the assisted human decisions are often no better than unassisted human decisions. A more extensive review of this literature is detailed by Lai et al. (2021). Of the works in this literature, our empirical approach is most related to that of Green and Chen (2019), which posits ways human decision-makers should respond to algorithms and tests these hypotheses in the lab. Relative to this literature, we focus on designing complementary recommendations, rather than focusing on describing ways in which humans misuse algorithms.

We join a collection of works spanning the areas of operations, economics, and computer science that design algorithms to assist humans in making better decisions. Bansal et al. (2019) and Donahue, Chouldechova, and Kenthapadi (2022) propose improvements to collaborative decision processes based on the private information of each agent. Balakrishnan, Ferreira, and Tong (2022) and Orfanoudaki et al. (2022) find that humans often incorrectly assume linear relationships and

then develop strategies to help human–algorithm decision systems overcome this issue. Athey, Bryan, and Gans (2020) and McLaughlin and Spiess (2022) model how effort costs and reference points, respectively, affect a human's adoption of algorithmic assistance. Snyder, Keppler, and Leider (2023) shows how workers' compliance with an algorithm's recommendation in a service system varies with the system's utilization. Baek et al. (2023) and Bastani, Bastani, and Sinchaisri (2022) apply ideas from reinforcement learning to incorporate past human actions into the recommendations of assistive algorithms. Caro and de Tejada Cuenca (2023) and Sun et al. (2022) use data on workers' deviations from algorithm advice to increase compliance. Noti and Chen (2023) focuses on when to provide assistance, and shows that strategically restricting advice can preserve the complementary strengths of the algorithm and human. Our approach differs from these prior works as it provides a causal framework to capture how a human's response varies with the underlying algorithm. Similar to the work of Ibrahim, Kim, and Tong (2021) (where human predictors assist algorithmic decision-makers), our results highlight the importance of private information in designing impactful collaboration.

Our approach to algorithmic recommendations relates to approaches in causal inference and algorithmic decision-making. First, we build upon policy learning and statistical treatment choice (Manski, 2004; Kitagawa and Tetenov, 2018; Athey and Wager, 2021). Second, we relate to instrumental-variables approaches in cases with imperfect compliance (Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996; Vytlacil, 2002), including the design of treatment rules with an instrument (Cui and Tchetgen Tchetgen, 2021; Qiu et al., 2021; Pu and Zhang, 2021). Third, we connect to marginal treatment effects (MTE) and the related approaches to policy-relevant treatment effects (Heckman and Vytlacil, 2005; Mogstad, Santos, and Torgovitsky, 2018; Mogstad and Torgovitsky, 2018). Fourth, we connect to the literature on selective labels and learning from human decisions (Lakkaraju et al., 2017; Kleinberg et al., 2017; Rambachan, 2024). Fifth, we join work by Imai et al. (2021) and Ben-Michael et al. (2025) in leveraging potential outcomes for a causal analysis of algorithm-assisted human decisions.

2 A Causal Model of Algorithmically Assisted Decision-Making

We introduce a model of joint human–algorithm decision-making, which builds upon our earlier work in McLaughlin and Spiess (2022). In this model, the designer chooses an algorithm that provides recommendations to a human decision-maker. The decision-maker observes each recommendation and subsequently makes a binary treatment choice. We model outcomes of this choice using the standard potential-outcomes framework from causal inference.

2.1 Treatment Choice and the Automation Benchmark

We consider the problem of taking a binary decision D, after which an outcome Y is realized. We capture the impact of this decision through a utility function $\mathcal{U}$, which describes the (scalar) evaluation $\mathcal{U}(Y)$ of obtaining the outcome Y achieved from implementing the decision D. Since we are ultimately interested in maximizing expected utility $\mathbb{E}[\mathcal{U}(Y)]$, we can assume without loss of generality that Y is scalar and that $\mathcal{U}$ is the identity, so that the outcome of interest is already measured in utility terms and the optimization target is $\mathbb{E}[Y]$.[1] For each problem instance, we assume that there are baseline characteristics X and U that have predictive power for the values Y takes for different choices of the treatment D.

We employ standard potential-outcomes notation from causal inference in order to express the effect of taking the decision $D \in \{1, 0\}$ on the outcome Y. Specifically, we assume that there are potential outcomes $Y(1), Y(0)$ representing the consequence of making the choices D=1 and D=0, respectively. For the actual decision D, the outcome is then realized as $Y = Y(D)$[2] . The potential outcomes $Y(1), Y(0)$ along with the covariates X and U follow a joint distribution $\mathbb{P}$.

The causal effect of treating a specific instance is $Y(1) - Y(0)$, and an optimal choice would be to choose $D = 1$ if and only if this difference is non-negative. However, we assume that the potential outcomes may be ex-ante unknown. If we knew both the features X available to the algorithm and the human information U, then the optimal (oracle) decision that maximizes expected utility would therefore be given by $D^* = \mathbf{1}[\tau^*(X, U) \geq 0]$, where $\tau^*(X, U) = \mathbb{E}[Y(1) - Y(0)|X, U]$ is the treatment effect conditional on both X and U. Since U is not generally available to the algorithm, a standard policy learning approach would be to instead estimate the conditional average treatment effect (CATE)

$$\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$$

and implement the rule

$$f^{\text{auto}} \in \underset{\pi: X \mapsto \{0,1\}}{\arg\max}\ \mathbb{E}[Y(\pi(X))], \qquad f^{\text{auto}}(x) = \mathbf{1}[\tau(x) \geq 0],$$

corresponding to the optimal direct choice by an algorithm that automates the decision. We refer to f^{auto} as the *automation benchmark*; it is the rule that standard policy learning targets.

Example 1 (Binary classification and hiring)**.** *As an example, consider a case where choosing treatment (D=1) comes with a cost of $c \in (0, 1)$ and leads to a binary benefit $Y^* \in \{1, 0\}$, relative to a net-zero outside option for choosing control (D=0). For example, D could represent a decision of whether to hire (D=1) or not hire (D=0) an applicant at cost c with latent ability Y^* that is good*

[1] For a general utility function $\mathcal{U}$, all results below apply to the potential outcomes $\mathcal{U}(Y(1))$ and $\mathcal{U}(Y(0))$ in place of $Y(1)$ and $Y(0)$. We therefore drop $\mathcal{U}$ from the notation.

[2] Throughout, the Stable Unit Treatment Value Assumption (SUTVA) that the outcome for one instance only depends on its own treatment is implicit to our framework.

(Y^*=1) or bad (Y^*=0). Here, we assume that the latent outcome Y^* is unaffected by treatment and unknown at decision time. The potential outcomes $Y(1), Y(0)$ associated with an instance with latent outcome Y^* are then the net benefits

$$Y(1) = Y^* - c, \qquad\qquad Y(0) = 0.$$

The optimal choice of treatment in this case compares expected benefit to cost, $D^* = \mathbf{1}[\mu^*(X,U) \geq c]$ for $\mu^*(X,U) = \mathbb{E}[Y^*|X,U]$. It depends only on the oracle prediction $\mu^*(X,U)$ of the (latent) outcome Y^*, and thus takes the form of a prediction policy problem *(Kleinberg et al., 2015). This treatment choice can also be seen as a binary classification problem. Specifically, maximizing expected benefit is equivalent to minimizing the expected classification loss $\mathbb{E}[\ell(Y^*, D)]$ with the loss function*

$$\ell(Y^*, D) = c_I \, \mathbf{1}[Y^* = 1, D = 0] + c_{II} \, \mathbf{1}[Y^* = 0, D = 1]$$

for costs $c_I = 1 - c, c_{II} = c$. Here, the type-I error cost $c_I > 0$ captures the relative cost of choosing $D = 0$ for $Y^* = 1$ (e.g., the company not hiring a good worker). The type-II error cost $c_{II} > 0$ captures the relative cost of choosing $D = 1$ for $Y^* = 0$ (hiring a bad worker). The choice that minimizes expected loss is then $D^* = \mathbf{1}\left[\mu^*(X,U) \geq \frac{c_{II}}{c_I + c_{II}}\right] = \mathbf{1}\left[\mu^*(X,U) \geq c\right]$ for $\mu^*(X,U) = \mathbb{P}(Y^*{=}1|X,U) = \mathbb{E}[Y^*|X,U]$, as before.

2.2 Algorithmic Assistance with Endogenous Response

Above, we have introduced the problem of assigning binary treatment D based on information X, U. We now assume that the decision D is taken by a human decision-maker with limited information and potentially misaligned preferences. Under assistance, the designer chooses a recommendation rule

$$f : X \mapsto R \in \{0, \varnothing, 1\},$$

where $R = 0$ and $R = 1$ recommend the two treatments and $R = \varnothing$ stands for no recommendation, and the human decision-maker observes R and takes the decision D. The objective $\max \mathbb{E}[Y(D)]$ is unchanged. What changes is that the designer controls the decision only through the recommendation. Specifically, we assume the decision D is taken by a decision-maker who (1) does not observe the characteristics X directly, (2) may observe private information determined by U, (3) may not share the designer's objective. We allow for arbitrary overlap between the information in X and the information observed by the decision-maker, so the model allows for none, some, or even all of the characteristics X to be available to the decision-maker. In causal terms, U is an unobserved confounder; it enters both the decision and the potential outcomes.

To improve the choices by the decision-maker, the designer implements an algorithm f that

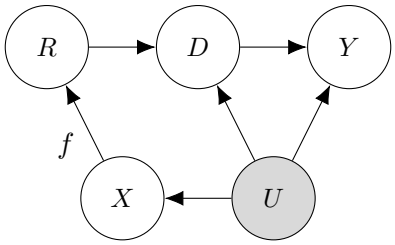


Figure 1: This diagram illustrates how the elements of our model are related to each other. The algorithm's information X determines the recommendation R through the rule f. The decision D responds to the recommendation and to the private information U of the decision-maker. The outcome Y depends on the decision and on U. The shading of U illustrates that the decision-maker's private information is never observable to the algorithm directly.

provides recommendations $R = f(X)$ to the decision-maker.[3] Thus, similar to the information design literature (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019), the designer (as the sender) has partial control over the information presented to the decision-maker (as the receiver) before the latter makes the final decision. We think of the algorithm as provided by an AI system that may learn from training data, but we abstract from the training process and focus on the resulting recommendation rule f itself.

Example 1 (continuing from p. 6)**.** *In the hiring example, we assume the designer is a company that implements an algorithmic recommendation system f that uses the characteristics X on workers to generate hiring recommendations $R = f(X)$, which influence the hiring decisions of the decision-maker, who is a manager at the company. The recommendations reduce the unstructured applicant data X to a simpler form that may still be informative about the worker's ability Y^*. For simplicity, we limit the algorithm to either recommend a decision to the manager (not hire, $R = 0$, or hire, $R = 1$) or not make a recommendation at all ($R = \varnothing$). In practice, recommendations may present more granular predictive information, such as a score that represents a scaled probability that a worker is good ($Y^* = 1$) and allows the managers to understand the algorithm's relative ranking of different workers. It may also include an explanation of the score, highlight pieces of X the algorithm deems particularly relevant, or include certainty information about the confidence of its prediction (Raghavan et al., 2020).*

The manager responds to the recommendation by making a hiring decision D which may depend on the recommendation R along with their private information U and unknown private preferences. This information U consists of all knowledge the manager has of the applicant, including potentially some of the information contained in X (such as the candidate's resume) as well as new information the manager learns through their own conversations with the applicant (such as prior relationships with employees at the company or interest in areas the company is looking to expand). Thus,

[3] We take f to be deterministic for simplicity. A natural extension considers randomized rules $R = f(X, \xi)$, with noise ξ independent of the instance.

we allow arbitrary correlation between X and U, but do not allow the manager to observe X directly. This effectively captures the idea that X and U contain shared information, but that neither necessarily dominates the other. Additionally, the manager may exhibit preferences that do not directly align with maximizing the outcome Y, preferring to hire individuals with whom they get along well (or have prior relationships) even when those applicants are less likely to be good than others. This flexibility allows us to capture the complex decisions managers exhibit empirically. We summarize the relationship between the elements of our model in Figure 1.

Throughout, we include the option that the algorithm does not make a recommendation, $R = \varnothing$, forcing the decision-maker to make an active decision. We believe that this design feature has formal advantages for modeling recommendations and practical relevance for achieving human–algorithm complementarity. First, we can consider unassisted decisions, those taken without any recommendation algorithm, as equivalent to those taken under a recommendation algorithm which always sends $R = \varnothing$, making unassisted decisions a special case of our framework. Second, building upon our earlier work in McLaughlin and Spiess (2022), forcing active decisions by avoiding recommendations in specific cases may reduce distortions in cases where human decision-makers know better. These beliefs are further supported by the work of Noti and Chen (2023), which finds that partially withholding algorithmic assistance (i) increases humans' reliance on advice when given and (ii) can be leveraged strategically to improve humans' prediction accuracy.

3 Making Sense of Recommendations with Potential Decisions

Having set up a treatment choice problem with endogenous responses for thinking about the design of recommendation algorithms, we now expand the potential-outcomes framework from causal inference to model how human decisions vary with recommendations. Viewing the decision-maker's decision in this way suggests that the descriptive and prescriptive analysis of recommendations shares common features with the investigation of instrumental variables in inference problems. We strengthen this connection by applying a monotonicity assumption similar to that of Imbens and Angrist (1994), which allows us to represent the remaining potential decisions in terms of the active decision the decision-maker makes when no recommendation is given and whether the decision-maker complies with the recommendation. This leads to a decomposition of the recommendation algorithm's objective based on the decision-maker's compliance choice. The objective captures the performance of the recommended action if the decision-maker complies and the performance of the decision-maker's active decision otherwise.

3.1 Recommendations Induce Potential Decisions

We now interpret the decision-maker's decision in terms of the potential-outcomes framework of Rubin (1974). Specifically, we write $D_f(r)$ for the "potential decisions" that the decision-maker

would take when receiving the recommendation r from the recommendation algorithm f when the decision-maker's information is U. Thus, the decision-maker's decisions can be seen as the random variable $D = D_f(R)$, where $R = f(X)$. The subscript f records that the response to a given recommendation may depend on the rule that produces it; we return to this dependence in Section 3.4.

Since the decision-maker's information U, the recommendation R, and the algorithm f are sufficient information to determine D, and the designer controls the recommendation algorithm f (and through it R), the decision is the potential decision $D_f(r)$ for each of the possible values $r \in \{0, \varnothing, 1\}$ of the recommendation. Since decisions are themselves binary and the recommendations take three values, we can enumerate these potential decisions into eight triplets which we categorize into four response types based on how the realized recommendation impacts the assisted decision-maker's decisions. These response types are similar in nature to the types developed by Angrist, Imbens, and Rubin (1996) for the analysis of instrumental variables.

Table 1 enumerates the potential choices associated with a decision-maker's response to recommendations and provides the aforementioned categorization. If the decision-maker maintains the same decision regardless of the recommendation, they ignore the recommendation (similar to always-takers and never-takers in causal inference). If they adopt the recommendation regardless of whether the algorithm recommends not hiring (0) or hiring (1) the applicant, then they comply with the recommendation. If, instead, the decision-maker always takes the decision that the algorithm does *not* recommend, then they defy the recommendation. If none of these cases hold, then it must be the case that the decision-maker will change their decision in response to the algorithm sending a recommendation, taking one decision when the algorithm sends no recommendation and taking the opposite decision if either recommendation is sent. While we maintain all these response types for now, we will later simplify the analysis by ruling out the defy and change types, analogous to the monotonicity assumption of Angrist, Imbens, and Rubin (1996).

3.2 Recommendations as Instruments

In the language of causal inference, our framework is an instrumental-variables problem: the recommendation is an instrument that affects the outcome only through its impact on the treatment, here the decision. In our hiring example, the algorithm's recommendations (corresponding to the instrument) impact the company's utility (playing the role of the outcome) only through managers' decisions (treatment). Rather than directly optimizing for the recommendation, the designer must understand how the decision-maker's decisions D vary in response to receiving different recommendations from different algorithms.

The condition that makes the recommendation an instrument in the sense of Angrist, Imbens, and Rubin (1996) is already built into the notation of Section 2, where the potential outcomes are indexed by the decision alone; we state it as an assumption to make visible what the notation rules

Response type	$D_f(0)$	$D_f(\varnothing)$	$D_f(1)$
Ignore	0	0	0
	1	1	1
Comply	0	0	1
	0	1	1
Defy	1	0	0
	1	1	0
Change	1	0	1
	0	1	0

Table 1: Enumeration and categorization of the potential decisions of the decision-maker in response to different realizations of the recommendation. The defy and change response types will later be ruled out by Assumption 2.

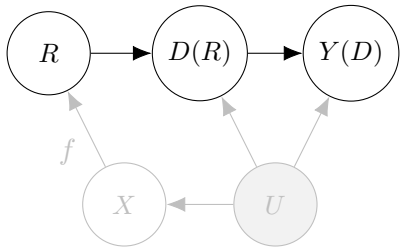


Figure 2: The recommendation as an instrument. The recommendation R affects the outcome only through the decision (Assumption 1) and is assigned as a function of the algorithm's information X; the private information U confounds the decision and the outcome.

out.

Assumption 1 (Exclusion)**.** *For every recommendation rule $f \in \mathcal{F}$, every decision $d \in \{0,1\}$, and every recommendation $r \in \{0,\varnothing,1\}$, the potential outcome of the decision does not depend on the recommendation: writing $Y_f(d,r)$ for the outcome that would be realized under rule f, recommendation r, and decision d, we have $Y_f(d,r) = Y(d)$.*

Exclusion fails when the recommendation changes how a decision is carried out, or when the human's action has dimensions that the binary D does not record, such as effort or timing. Since $R = f(X)$, the recommendation is independent of $(Y(1), Y(0), D_f(\cdot))$ conditional on X but not unconditionally, so all instrumental-variables statements in this paper are conditional on $X = x$, with U in the role of the unobserved confounder of D and Y. What is *not* assumed is that the potential decisions $D_f(\cdot)$ are the same across rules f; Section 3.4 returns to this point. Figure 2 shows the structure.

3.3 Monotonic Responses, Compliance, and Active Decisions

Above, we have proposed a potential-outcomes framework for human decisions in response to recommendations. We now introduce assumptions that limit the potential decision to simplify their analysis. We then use these simplifications to shed further light on the structure of the optimal recommendation algorithm in terms of compliance and active decisions.

In the context of the four response types introduced in Section 3.1 and listed in Table 1, behavior consistent with the ignore and comply responses is easier to make sense of than behavior consistent with the defy and change responses. If the manager believes the algorithm made a bad recommendation given their private information U about the applicant, they are likely to ignore the recommendation. Conversely, if they believe the algorithm makes a good recommendation given their private information U about the applicant, they will comply. Using recommendations as a threshold requirement for considering an applicant, as done by hiring managers interviewed by Li et al. (2021), also generates potential outcomes consistent with only ignore and comply responses. On the other hand, defy and change behaviors are counter-intuitive and typically need bad recommendation algorithm designs, perverse correlation between X and U, or adversarial managerial incentives to be justified rationally. Consequently, we now make an assumption to eliminate the potential outcomes associated with defy and change responses, similar to the monotonicity assumption of Angrist, Imbens, and Rubin (1996) in the context of instrumental variables and its extensions to multivariate discrete treatments (including Lee and Salanié, 2018).

Assumption 2 (Monotonic response to R)**.** *Let $\preceq$ be the ordering of decisions with $0 \preceq 1$. We assume, for all $f \in \mathcal{F}$, that $D_f(\cdot)$ is such that $D_f(0) \preceq D_f(\varnothing) \preceq D_f(1)$.*

At a high level, this assumption restricts the direction of the decision-maker's response to follow the algorithm's recommendation. In the hiring example, hire recommendations ($R = 1$) can turn not-hire decisions into hire decisions ($D = 0 \to D = 1$) but cannot turn hire decisions into not-hire decisions ($D = 1 \not\to D = 0$). Not-hire recommendations ($R = 0$) do the exact opposite.

The monotonicity restriction exactly rules out the bottom four rows of Table 1. It can be considered a restriction on plausible recommendation rules $\mathcal{F}$. Specifically, it rules out that recommendation rules are used that would make sense to defy (such as a rule that suggests hiring bad candidates) or that would make sense to observe the change behavior for (such as a rule that only gives a recommendation for bad candidates and never gives one for good candidates).

When the decision-maker's behavior is limited by Assumption 2, we can fully describe their behavior for some specific recommendation rule f in terms of their active decisions $A = D_f(\varnothing)$ which they exhibit in response to the algorithm f being implemented, along with their compliance behavior C:

Proposition 1 (Simplified potential outcomes)**.** *Under Assumption 2, the potential outcome triplet $(D_f(0), D_f(\varnothing), D_f(1))$ can be fully expressed by the active decision $A_f = D_f(\varnothing) \in \{0, 1\}$, which*

C_f	$D_f(0)$	$D_f(\varnothing)$	$D_f(1)$
0	A_f	A_f	A_f
1	0	A_f	1

Table 2: Reduction of potential outcomes under Assumption 2. This organization acts as a visual representation of Proposition 1, showing how the values that the potential outcomes take reduce to the active decision $A \in \{0, 1\}$ absent a recommendation and the compliance $C \in \{0, 1\}$.

expresses the decision-maker's decision when not given a recommendation, and the compliance type $C_f \in \{0, 1\}$*, which specifies whether the decision-maker adopts* $R = f(X)$ *when the recommendation is presented.*

Table 2 visualizes Proposition 1 by restating the remaining potential-decision types in terms of (A, C) once Assumption 2 is applied. The only information determining the decision D outside of the recommendation r is the active decision A and whether the response type C of the decision-maker is 0 (ignore) or 1 (comply). The proposition follows directly from an inspection of Table 1. For completeness, a formal proof is given in Appendix A. Written out, the representation is

$$D = \begin{cases} R & \text{if } R \neq \varnothing \text{ and } C_f = 1, \\ A_f & \text{otherwise,} \end{cases}$$

which is the identity that the decomposition of the objective in Section 3.5 uses.

3.4 Dependence on the Rule

The potential decisions $D_f(r)$ depend not only on possible recommendation values r of R, they also depend on the whole mapping f from characteristics X to recommendations $R = f(X)$. This is because how a *specific* recommendation is interpreted depends on how recommendations are given overall. For example, if an algorithm recommends hiring for *everybody*, then it may change individual decisions differently from it recommending to hire only a few applicants. This is related to the concept of *interference* in causal inference, where potential outcomes may be functions not just of their own treatment assignment, but also of those of others. Potential decisions may also depend on context: on how the rule is explained, on the interface, or on the decision-maker's experience with the algorithm. The rule f can then be read as the design as a whole. This dependence highlights the challenge in estimating good recommendation algorithms under minimal assumptions or using limited data (which we return to in Section 5). Effective ways of forming recommendation algorithms will, therefore, rely on assumptions that limit the complexity of potential decisions,

including their dependence on the overall recommendation algorithm.[4]

3.5 Objectives of Recommendation Algorithms

Having introduced a framework to capture potential decisions in response to recommendations, we now consider the optimal recommendation algorithm as the solution to an optimization problem that maximizes the expected outcome achieved by the decisions the algorithm induces. We show that the recommendation's objective can be separated into a triage term that assumes that recommendations are implemented directly and a term that captures the decision-maker's endogenous response to the recommendation. We compare this optimal recommendation algorithm to solutions that instead directly target optimal decisions, and provide an intuition for why the automation benchmark will rarely be the optimal recommendation rule.

In general, the optimal recommendation algorithm is the one that maximizes the expected outcome of the decision taken by the human decision-maker after seeing the recommendation,

$$f^* = \underset{f:X\mapsto R}{\operatorname{argmax}}\ \mathbb{E}\left[Y\right] \text{ where } R = f(X), D = D_f(R), Y = Y(D). \tag{1}$$

In our hiring example, the optimal recommendation algorithm for the company to implement is the mapping from applicant worker data to recommendations that leads to good hiring decisions by the manager. As this example illustrates, this optimization problem is generally quite difficult to solve, given the uncertainty about how managers will respond to recommendations. As a result, designers often implement algorithms that instead recommend the optimal decision based on the information available to the algorithm, the automation benchmark of Section 2,

$$f^{\text{auto}} = \underset{f:X\mapsto R}{\operatorname{argmax}}\ \mathbb{E}\left[Y(R)\right] \text{ where } R = f(X), R \neq \varnothing, \tag{2}$$

implicitly assuming that the recommendation $R = f(X)$ gets implemented directly ($D = R$) and ignoring the algorithm's ability to not send a recommendation ($R = \varnothing$). An extension suggested in previous work by Raghu et al. (2019) is to triage instances into those where algorithms outperform humans and those where humans outperform algorithms. We write $\underline{D} = D_{f^\varnothing}(\varnothing)$ where $f^\varnothing \equiv \varnothing$ for the decision taken by the decision-maker in the absence of a recommendation algorithm. If we

[4]Without additional assumptions, this interference challenge implies, in our case, that the potential decision function $D_f(\cdot)$ for a *specific* instance (X, U) is only ever evaluated at its implied recommendation $R = f(X)$. However, the decision-maker does not observe X, so X affects D only through R: for a given rule f, the potential decision function $D_f(\cdot)$ is a function of U alone. This exclusion restriction at the decision stage is what makes $D_f(r)$ well defined for recommendations $r \neq f(X)$, and the notation then also captures variation in recommendations given for instances that are otherwise equivalent from the perspective of the human decision-maker. That is, for a given instance U, they capture how decisions vary across possible realizations of $f(X)$ stemming from variation in X. This formalization, therefore, implies that two instances that present with the same U, R values are not distinguishable to the decision-maker, leading to the same decision even if their X varies. (Here, any noise in decisions would be captured through U.)

assume that the decision-maker takes this decision when not given a recommendation ($f(X) = \varnothing$) and that otherwise the recommendation is implemented directly, then this objective and its solution correspond in our framework to

$$f^{\text{triage}} = \underset{f:X\mapsto R}{\text{argmax}}\ \mathbb{E}\left[Y(R)\,\mathbf{1}\left[R \neq \varnothing\right] + Y(\underline{D})\,\mathbf{1}\left[R = \varnothing\right]\right] \text{ where } R = f(X). \tag{3}$$

In practice, recommendations may not be followed perfectly. In addition, even the decision for $R = \varnothing$ may itself be different from the decision $\underline{D}$ without a recommendation algorithm, and we may specifically hope it to improve since additional information and effort may be available for those cases (Raghu et al., 2019; Athey, Bryan, and Gans, 2020).

We now compare these different objectives and solutions, and show that the optimal recommendation (f^*) differs from triage (f^{triage}) in that it also incorporates the compliance behavior of the decision-maker and considers how active decisions may differ from unassisted decisions. Specifically, the objective that defines the optimal recommendation can be decomposed into the objective of the triage problem (which we call the triage effect, TE) and the response effect (RE) from decision-makers deviating from the triage decision,

$$\begin{aligned} f^* = \underset{f:X\mapsto R}{\text{argmax}}\ & \overbrace{\mathbb{E}\left[Y(R)\,\mathbf{1}\left[R \neq \varnothing\right] + Y(\underline{D})\,\mathbf{1}\left[R = \varnothing\right]\right]}^{TE\text{ – Triage Effect}} \\ & + \underbrace{\mathbb{E}\left[\left(Y(D) - Y(R)\right)\,\mathbf{1}\left[R \neq \varnothing\right] + \left(Y(D) - Y(\underline{D})\right)\,\mathbf{1}\left[R = \varnothing\right]\right]}_{RE\text{ – Response Effect}} \\ & \text{where } R = f(X), D = D_f(R), Y = Y(D). \end{aligned} \tag{4}$$

In addition to the triage effect from (3), this decomposition has a second part representing deviations from recommendations and unassisted decisions. It can represent both improvements (in cases where the human decision-maker uses their private information to override a suboptimal recommendation or adjusts their active decision based on the knowledge they receive through the recommendation algorithm) as well as mistakes (when the decision-maker goes against a correct recommendation or takes an active decision that is worse than in the case without any recommendations).

In the language of instrumental variables, the left-hand side of (4), $\mathbb{E}[Y(D_f(R))]$, is the intent-to-treat value of the rule: the recommendation is assigned, and the human decides.

The two components also highlight the challenge of learning good recommendations from data. The triage effect (TE) is straightforward to estimate, provided the designer has access to historical data through which they can learn the joint distribution of characteristics X, potential outcomes $(Y(1), Y(0))$, and unassisted decisions $\underline{D}$.[5] On the other hand, the response effect depends on the

[5]This is the case in the experiment of Section 4, where the latent ability of every applicant is known. With only the realized outcome $Y = Y(\underline{D})$ in the data, the triage effect is not identified without further assumptions; Section 5

potential decisions of the decision-maker, which complicates its evaluation for any f under which decision data has not already been collected. Understanding $D_f(r)|X, Y(1), Y(0)$ may require more than simple experimental evaluation based on local changes in recommendations. These decision distributions may vary arbitrarily with changes to f. Solving for the optimal recommendation then requires understanding this distribution across a set of reasonable recommendation algorithms $\mathcal{F} \subseteq \{f; f : X \mapsto R\}$. This makes designing an optimal recommendation algorithm an intractable task without additional assumptions (even when we just want to robustly outperform the unassisted decisions $\underline{D}$).

As an immediate consequence of the simple representation of response types in Proposition 1, we can decompose the objective of the recommendation algorithm in terms of compliance:

$$f^* = \underset{f:X\mapsto R}{\operatorname{argmax}} \mathbb{E}\left[Y(R)\,\mathbf{1}\left[C = 1 \text{ and } R \neq \varnothing\right] + Y(A)\,\mathbf{1}\left[C = 0 \text{ or } R = \varnothing\right]\right] \quad \text{where } R = f(X) \tag{5}$$

Similarly, this explicit notation allows us to shed further light on the response effect from (4) in terms of compliance and deviations between active decisions and unassisted decisions:

$$RE = \mathbb{E}\left[(Y(A) - Y(R))\;\mathbf{1}\left[R \neq \varnothing, C = 0\right] + (Y(A) - Y(\underline{D}))\;\mathbf{1}\left[R = \varnothing\right]\right]$$

The first term in this decomposition expresses deviations from recommendations when they are ignored, while the second term captures the change in behavior for those instances where no recommendation is given that comes from the presence of the recommendation algorithm.

The same representation gives the objective in terms of treatment effects. With $\tau = Y(1) - Y(0)$, Proposition 1 implies $Y(D) = Y(A_f) + (Y(R) - Y(A_f))\,C_f\,\mathbf{1}[R \neq \varnothing]$, and since $Y(1) - Y(A_f) = (1 - A_f)\tau$ and $Y(0) - Y(A_f) = -A_f\tau$,

$$\mathbb{E}[Y(D_f(R))] = \mathbb{E}[Y(A_f)] + \mathbb{E}\big[\tau\,(1 - A_f)\,C_f\,\mathbf{1}[R = 1]\big] - \mathbb{E}\big[\tau\,A_f\,C_f\,\mathbf{1}[R = 0]\big]. \tag{6}$$

The value of a rule is the value of the active decisions, plus the treatment effects of the instances that a hire recommendation moves into treatment, minus the treatment effects of the instances that a not-hire recommendation moves out of it. What matters for the recommendation at x is therefore the average effect among the compliers whose decision it changes, $\mathbb{E}[\tau \mid X = x, A_f = 0, C_f = 1]$ for a hire recommendation and $\mathbb{E}[\tau \mid X = x, A_f = 1, C_f = 1]$ for a not-hire recommendation, and not the conditional average treatment effect $\tau(x)$ that defines the automation benchmark.

takes up this case.

3.6 Optimal Recommendations under Specific Assumptions

Even with the simplifications introduced by Assumption 2, it is not generally possible to learn the optimal recommendation algorithm from past data $(Y, X, \underline{D})$ on outcomes, characteristics, and unassisted human decisions. This is because the joint distributions of active decisions and compliance with outcomes and characteristics, (Y, X, A, C) cannot be identified from such limited data without further assumptions. However, we can now formulate conditions on these unobserved quantities to derive optimal learnable recommendation algorithms in specific cases, thus providing a formal analog and extension of the discussion in Section 3.5:

- If compliance is perfect ($C \equiv 1$) and decisions in instances without recommendations are not affected ($A \equiv \underline{D}$), then the optimal recommendation algorithm is the optimal triage solution f^{triage} (which can be learned from data on $(Y, X, \underline{D})$).

- If compliance is perfect ($C \equiv 1$), and recommendations are limited to be binary ($R \neq \varnothing$) or human decisions are dominated by recommended actions (such as when $\mathbb{E}[Y(A)|X] \leq \max_{r \in \{1,0\}} \mathbb{E}[Y(r)|X]$ for all $f \in \mathcal{F}$), then the optimal recommendation algorithm is the automation benchmark f^{auto} (which can be learned from data on (Y, X)). This applies, specifically, in the case with perfect compliance and no private information of the human decision-maker.

At the same time, more nuanced assumptions or data on compliance behavior may be necessary to achieve complementarity in practical cases. This may include the cases below:

- If the recommendation algorithm is independent of compliance and active decisions ($A \equiv \underline{D}$), then the optimal recommendation can generally be seen as the optimal triage decision *among compliers only*,

$$\underset{f:X \mapsto R}{\operatorname{argmax}} \, \mathbb{E}\left[Y(R)\,\mathbf{1}\left[R \neq \varnothing\right] + Y(\underline{D})\,\mathbf{1}\left[R = \varnothing\right] | C = 1\right] \text{ where } R = f(X).$$

  However, unlike the case of full compliance, we now have to learn about compliance behavior in addition to active decisions.

- Even if the decision-maker makes optimal decisions given their available information, we would require information beyond $(Y, X, \underline{D})$ since the optimal recommendation would also depend on the structure and information content of the private signal U.

In our experiment below, we combine information about the joint distribution of $(Y(1), Y(0), X, U)$ with assumptions on active decisions A and compliance C in order to solve for sensible recommendation algorithms.

4 Complementary Recommendations in an Experiment

We now apply our framework to the design and analysis of recommendation algorithms in a controlled trial. In a pre-registered online experiment, study subjects participate in an incentivized hiring game.[6] This empirical application showcases our framework's ability to map realistic assumptions about decision-maker behavior to the design of practical recommendation algorithms that realize complementarity between humans and algorithms.

In the experiment, we collect hypothetical hiring decisions while varying the structure of the recommendation algorithm at the subject level. We apply different assumptions on how subjects will behave within our framework to generate these recommendation algorithms. We find that subjects, on average, improve their decisions with algorithms whose recommendations are focused on providing suggestions that are complementary to the subjects' private information. Moreover, subjects given these focused recommendations perform better than subjects who see naive recommendations based on optimal algorithmic predictions and outperform decisions by the human or algorithm alone. In the notation of the framework, the treatment D is the hiring decision; the potential outcomes are $Y(1) = +1$ for a good applicant and $Y(1) = -1$ for a bad one, and $Y(0) = 0$, which is the hiring example of Section 2 with $c = 1/2$ after rescaling; X is the applicant's personality type, which only the algorithm sees, and U is the applicant's role, which only the subject sees, together with any noise in the subject's decisions. Each treatment arm is a rule f, randomized across subjects, so that each arm identifies the value $\mathbb{E}[Y(D_f(R))]$ of its rule, and the share of correct decisions equals $\mathbb{E}[Y(D)]$ up to the constant $\mathbb{P}(Y^* = 0)$. Because the latent ability Y^* is known for every applicant, both potential outcomes are known for every instance, which is the data condition that Section 5 calls observed outcomes.

4.1 Experimental Setup

We implement a hiring game as part of an online experiment to test subjects' performance when using recommendation algorithms in a setting where both the subject and algorithm have private information. Specifically, we label 25 hypothetical applicants as good or bad based on the role they are applying for and their personality type, and ask subjects to only hire good applicants. The role is only known to the human subject and the personality type is only known to the algorithm. We measure the change in the accuracy of subjects' hiring decisions in response to different recommendation algorithms (treatments) that vary at the subject level.

For each subject, the experiment proceeds as follows. Subjects arrive at a Qualtrics survey from the recruiting site Prolific and learn they will make hiring decisions on 25 hypothetical applicants, of which there are good and bad types (referred to as the applicant's ability). They learn that each applicant's ability Y^* is determined by the relationship of the applicant's personality type

[6] The pre-analysis plan for this experiment can be found at www.socialscienceregistry.org/trials/11857.

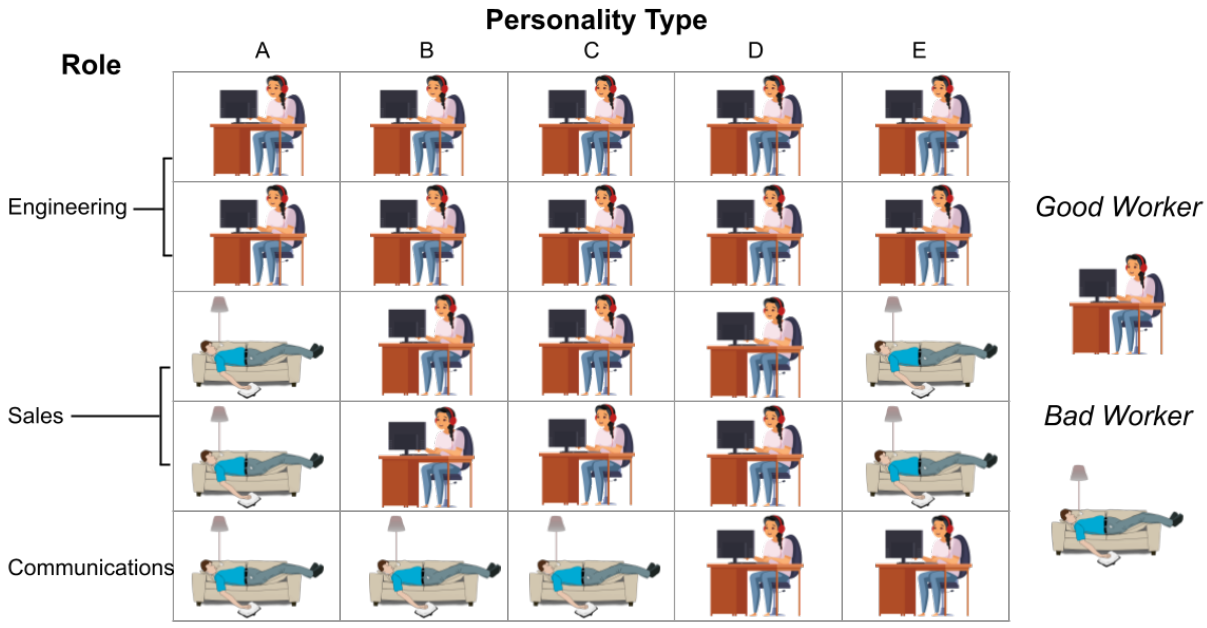


Figure 3: Summary of how the information available to the subject (the applicant's role) and the information of the algorithm (the applicant's personality type) relate to the applicant's ability (bad or good). This information is shown to all participants in the experiment and is easily accessible throughout their decision-making process.

X (denoted by letters A–E) to their intended role U (Engineering, Sales, or Communications), as shown in Figure 3. Then they answer three comprehension questions about how ability Y^* is determined and, conditional on answering correctly, we randomize them into a treatment.

Subjects receive recommendations from algorithms to assist them with making hiring decisions, except for those subjects in the control arm. Unlike the study subjects, these algorithms can access the applicants' personality types, but not their role.[7] Subjects then make 25 hiring decisions with access to the distribution of applicants as represented in Figure 3, the recommendations of their assigned algorithm, and an explanation of that specific recommendation algorithm. An example of a hiring decision is given in Figure 4. When making these hiring decisions, subjects know the applicant's role, but are unaware of the applicant's personality type beyond any information conveyed by the recommendation.

We incentivize subjects to hire good applicants and to not hire bad applicants. Specifically, as compensation for participating in our experiment, subjects receive \$2.04 for their time (estimated to be 12 minutes) and the chance to earn a \$2.00 bonus as a performance incentive. They earn this bonus if their hiring decision on a randomly selected applicant is correct (hire if good, not hire if

[7]Subjects then answer three more comprehension questions about the specific algorithm they will interact with, but continue in the experiment regardless of their answers. We leverage these additional questions for robustness checks below.

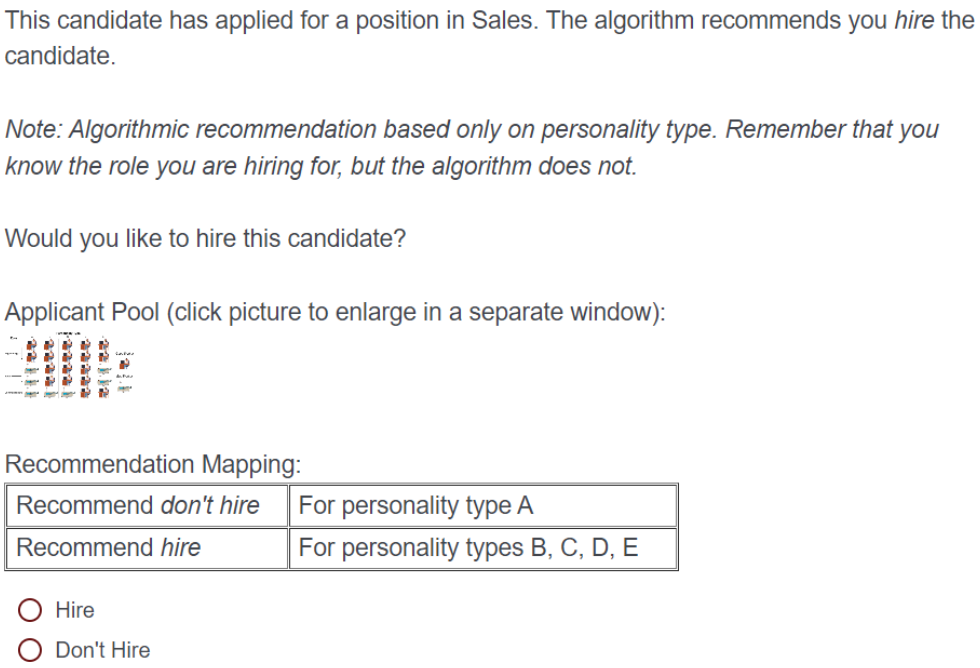


Figure 4: Sample hiring decision from the *Predictive* treatment.

bad). Note that this induces equal type I and type II losses ($c_I = c_{II}$) from the perspective of our subjects, and we analogously evaluate their performance using the percentage of correct decisions taken. Subjects see all 25 applicants in a random order without replacement to make their hiring decisions but don't learn whether they made correct decisions.

The joint distribution of applicant role, personality type, and ability (pictured in Figure 3) is designed to test subjects' willingness to comply with recommendations across applicant roles while allowing the recommendation to provide useful information in cases the applicant role is relatively uninformative. When subjects review an applicant for an Engineering role, they know the worker is good. Subjects are much more uncertain for Sales applicants (60% of which are good) and Communications applicants (40% of which are good). The algorithm's private information (personality type) varies significantly across the cases in which subjects are uncertain, thus containing useful information to help them make their decisions. Moreover, both the subject and algorithm's private information are of equal value in that a rational agent with access to either the role or the personality type would take the correct decision 76% of the time.

4.2 Treatments

The recommendation algorithm a subject receives represents the treatment in our experiment and is randomized at the subject level. In addition to four treatment arms representing different recommendation algorithms, there is a control arm in which no recommendation is given. These

Algorithm	A	B	C	D	E
Control	∅	∅	∅	∅	∅
Predictive	0	1	1	1	1
Complementary	0	1	1	1	0
Triage	0	∅	∅	1	1
Complementary Triage	0	∅	∅	1	0

Table 3: Recommendations that each algorithm sends across applicants' personality types $X \in \{\text{A}, \text{B}, \text{C}, \text{D}, \text{E}\}$. Note that personality types B and C are equivalent for the distribution of outcomes as well as in terms of recommendations given.

recommendation algorithms differ in the way they map personality types X to recommended actions $R = f(X) \in \{0, \varnothing, 1\}$, where the recommendations follow our earlier example and correspond to not hiring, no recommendation, and hiring the applicant, respectively. We now describe the four different algorithms and link them to the implicit assumptions about compliance and active decisions that motivate their design. Additionally, Table 3 lists the recommendations that these algorithms provide across personality types.

Control. Subjects in this condition do not observe a recommendation algorithm. We use this treatment to collect information on the distribution of subjects' unassisted decisions.

Predictive. The algorithm in this condition sends a hire recommendation if at least half of the applicants associated with the given personality type are good and sends a not-hire recommendation otherwise. This *Predictive* algorithm is the optimal decision an algorithm can take by itself to minimize overall errors, so it corresponds to f^{auto} in the framework of Section 3. It would correspond to the optimal *recommendation* if compliance were perfect and active decisions only led to worse hiring decisions.

Complementary. These subjects observe an algorithm that sends a hire recommendation if at least half of the Sales and Communications applicants associated with the given personality type are good and sends a not-hire recommendation otherwise. This algorithm ignores Engineering applicants on the assumption that subjects will perform effectively by themselves when asked whether to hire these applicants, and instead sends the optimal decision for the subpopulation of Sales and Communications applicants. It corresponds to the optimal recommendation if subjects do *not* comply with the recommendation for Engineering candidates, but comply perfectly for the remaining applicants.

Triage. These subjects observe a version of the *Predictive* algorithm that implements a 'safety check', causing the algorithm not to send recommendations for some personality types. Specifically, the *Triage* algorithm only sends the *Predictive* algorithm's recommendation of hire if at least half of the Communications applicants associated with the given personality type are

good and it only sends the *Predictive* algorithm's recommendation of not hire if at least half of the Engineering and Sales applicants associated with the given personality type are bad. As a consequence, this algorithm does not send recommendations for the B and C personality types (which are equivalent). The *Triage* algorithm is the optimal triage algorithm (f^{triage}) for a rational agent with the subjects' private information.

Complementary Triage. These subjects observe an algorithm that combines the logic of the *Triage* and *Complementary* algorithms. These subjects receive the *Complementary* algorithm's recommendation of hire if at least half of the Communications applicants associated with the given personality type are good and receive the *Complementary* algorithm's recommendation of not hire if at least half of the Sales applicants associated with the given personality type are bad. The *Complementary Triage* algorithm is the optimal algorithmic triage over the subpopulation of Sales and Communications applicants.

We summarize the logic behind these algorithms by examining their implicit assumptions about the subjects' compliance choices and active decisions. Specifically, Table 4 presents each algorithm as a solution to a specific combination of assumptions. For compliance choices, we distinguish between perfect compliance (that is, following the recommendation and only making an active decision when no recommendation is given) and selective compliance. By selective compliance, we mean that subjects always make an active decision for Engineering candidates and comply otherwise. For active decisions, we distinguish between completely random decisions and sophisticated decisions. Sophisticated decisions assume that the subject uses their private role information effectively when they take active decisions, that is, only hiring Engineering and Sales candidates, which is optimal absent recommendations and across each of our treatment arms. For example, if compliance is perfect and decisions are sophisticated, then we assume that subjects follow all recommendations and correctly decide in the absence of a recommendation when they do not receive one.

Using these combinations of behaviors, we can interpret the above treatment arms as optimal recommendation algorithms. First, the *Predictive* algorithm is optimal with perfect compliance and random active decisions. Second, the assumptions under which the *Triage* algorithm is optimal are that compliance is perfect and active decisions are sophisticated. In this case, the subject follows the recommendation when it is given and otherwise relies on their private information to decide. Third, the *Complementary* algorithm is optimal when subjects comply for non-Engineering candidates only, and make random decisions when no recommendation is sent. Fourth, the *Complementary Triage* recommendations achieve optimal outcomes when subjects do not comply for (and always hire) Engineering candidates, and effectively use their private information to make a sophisticated decision when they do not receive a recommendation. If the human were replaced by a rational agent, this treatment arm would achieve the best performance on average, with only one mistake (not hiring a type-E Communications candidate).

	Compliance C	
Active decision A	Perfect	Selective
Random	*Predictive*	*Complementary*
Sophisticated	*Triage*	*Complementary Triage*

Table 4: Optimal algorithms under different combinations of compliance choices and active decisions.

Although final recommendations only depend on the personality type X, we note that constructing these algorithms requires an understanding of the joint distribution of X with the ability Y^* as well as the role U. Whether such a joint distribution can be learned in practice depends on the type and quality of data that is available at the time of the construction of the algorithm. Here, we focus instead on an analysis of the relative performance of different recommendation algorithms in a setting where we (and the study subjects) know the joint distribution of (Y^*, X, U). In Section 5, we will return to the question of how to design recommendations when the distribution of U remains unknown and our only insight into the humans' behavior is through their unassisted decisions ($\underline{D}$).

4.3 Data and Main Results

Before analyzing the results of the experiment, we describe the data collection process, final sample sizes for each treatment, and basic performance statistics. Among 1675 surveys begun, about 35% failed to answer at least one of the initial comprehension questions and were screened out, leaving us with 1083 subjects that completed our study. Subjects were split approximately equally between all arms except *Control*, which received half as many observations as the other treatments. Specifically, each of the four treatment arms contained approximately 240 participants who, like the approximately 120 *Control* subjects who did not get recommendations, each took 25 hiring decisions for a total of around 6,000 decision responses per treatment and 3,000 for *Control*. Overall, the subjects performed well in the experiment and within the estimated time frame. The subjects of our experiment made the correct decision 78.1% of the time. The average duration of the experiment was around twelve minutes. However, both the performance (standard deviation of 13 percentage points) and duration (standard deviation of almost eight minutes) show significant heterogeneity across subjects.

We compare treatments by average subject performance, with the main results reported in Table 5. Subjects in the *Predictive* treatment, who received recommendations from an algorithm that was trained to predict outcomes, got around 76% of hiring decisions correct, significantly above the around 70% in the *Control* group that was not supplied with any recommendations. The *Triage*, *Complementary*, and *Complementary Triage* treatments were all designed to provide recommen-

Treatment arm	Optimal (%)	(SE)	Hire (%)	(SE)	Deviated (%)	(SE)
Control	69.5	(0.7)	67.0	(1.1)	N/A	
Predictive	76.2	(0.5)	75.3	(0.7)	19.7	(0.9)
Complementary	81.6	(0.7)	66.9	(0.6)	25.6	(0.8)
Triage	78.8	(0.8)	65.9	(0.6)	23.5	(1.0)
Complementary Triage	80.1	(1.1)	60.9	(0.8)	28.2	(1.0)

Table 5: For each treatment we list the percentage of subjects' decisions that were (i) optimal, (ii) hire, or (iii) deviating from the recommendation. Bootstrap standard errors are clustered at the subject level.

dations that complement human decisions, and subjects in these groups on average outperformed subjects in the *Predictive* treatment. Hence, the data confirms that the algorithm designed to take optimal decisions by itself (given by the *Predictive* treatment) is not the optimal recommendation algorithm in this case. Across the treatments, subjects assigned to the *Complementary* algorithm perform best, with over 81% optimal decisions (although this performance is statistically indistinguishable from the *Complementary Triage* subjects). Overall, the results suggest that subjects tend to make smart compliance choices and frequently (and correctly) override the algorithm for Engineering candidates. At the same time, decisions are inefficiently noisy.

We compare subject performance to useful reference points in Figure 5 (justified further in Wu et al., 2023). Throughout, we focus on the fraction of cases in which subjects make the optimal decision on an applicant (hire if good, not hire if bad). Within each treatment, we compare subject performance to the performance when following the algorithm perfectly (corresponding to the Triage Effect of Section 3.5), and plot the impact of subject deviations (Response Effect). We also provide the best achievable performance by an algorithm acting alone (to benchmark which treatments were able to outperform algorithmic automation) as well as the performance achievable by a rational agent acting with the subject's information (to act as a performance upper bound). The decomposition reveals that subject performance tends to improve by deviating from the algorithm (at least in cases other than the *Predictive* arm, in which case subject performance is equivalent to perfectly following the algorithm). Interestingly, machine-assisted human performance is better for the somewhat simpler *Complementary* algorithm than for the theoretically optimal *Complementary Triage*. Overall, we think that the patterns in Figure 5 are in line with good compliance choices, noisy active decisions, and a penalty for overly complex recommendation algorithms.

Our experiment documents complementarity between humans and algorithms when recommendations are designed well. Subjects in the *Triage*, *Complementary*, and *Complementary Triage* treatments all performed significantly better (at the 5% significance level) than humans by themselves (*Control*) or the algorithm by itself, thus achieving human–AI complementarity, while subjects in the *Predictive* treatment could only match the performance humans or algorithms could

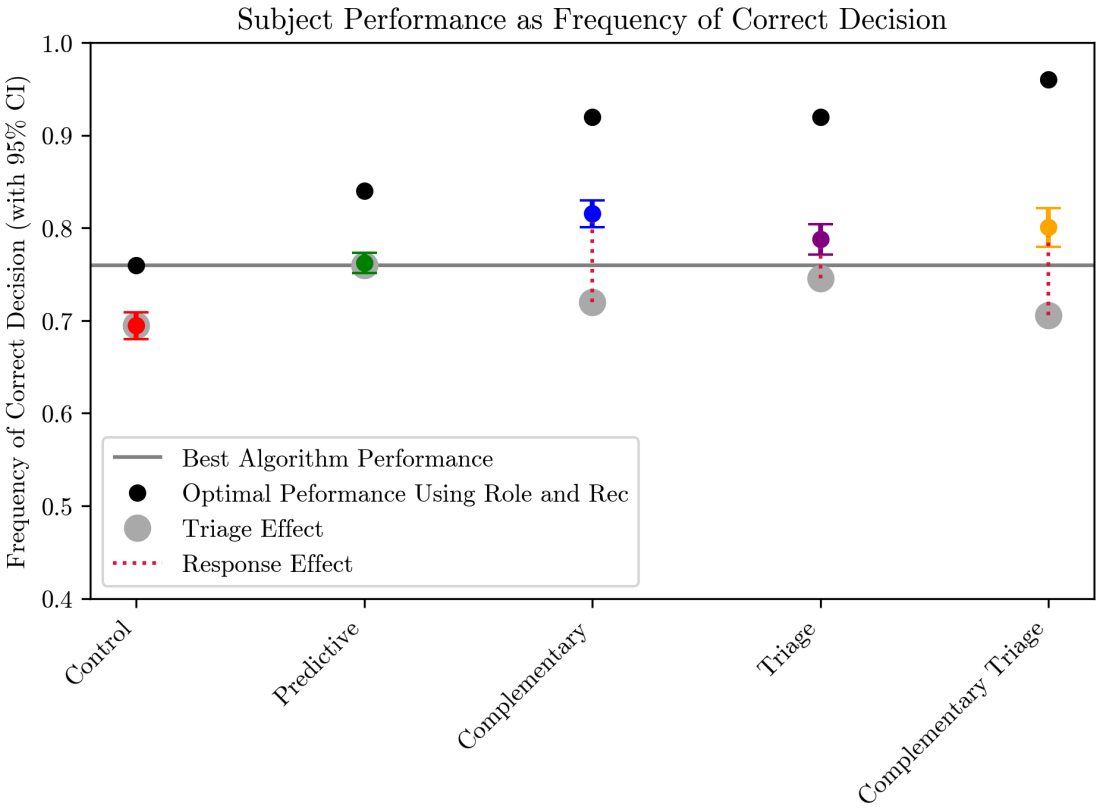


Figure 5: Average subject performance (through fractions of optimal hiring decisions) across treatments. 95% bootstrap confidence intervals are clustered at the subject level. In addition, the triage performance (where compliance is perfect and not sending a recommendation uses the decision of *Control*) is given for each treatment by a gray dot, while the Response Effect from Section 3.5 is shown by the vertical red dashed lines. The optimal algorithm performance is shown by the gray horizontal line. The optimal performance achievable by a rational agent with the subject's knowledge (role and algorithmic recommendation) is given for each treatment arm as a black dot.

(theoretically) achieve by themselves.

4.4 Additional Analysis and Robustness Checks

To understand performance differences across treatment arms further, we separate subjects' propensity to hire applicants across roles and personality types in Figure 6, and relate this behavior to the recommendations subjects receive. From the figure, we can identify that subjects' unassisted decisions mostly match the average quality of candidates (which would be in line with a probability-matching model where subjects match the probability they hire an applicant with the probability they are good). Their compliance choices are generally sophisticated (in that they correctly override the algorithm for Engineering candidates), although somewhat random. Note that our framework predicts that the *Complementary* algorithm is optimal if subjects make somewhat noisy active decisions but make selective compliance choices. This is in line with our empirical findings, where the *Complementary* algorithm is the best among those algorithms we tested and likely the optimal recommendation algorithm we could have implemented.

As an additional robustness check for our results in this section, we exploit the additional comprehension questions we asked subjects about the recommendation algorithm they used. Specifically, 376 of the 961 study subjects who were given a recommendation did not answer one of the comprehension questions correctly about their specific algorithm. We reproduce all analysis on both the sub-population of subjects who answered all recommendation comprehension questions correctly and those who answered at least one question wrong in Appendix B. Our results strengthen on the sub-population who answered all comprehension questions correctly, while, as expected, results are very noisy and largely insignificant on the sub-population of subjects who answered at least one question wrong. These results support conclusions by Bansal et al. (2019) that a subject's understanding of an algorithm is vital to achieving complementarity.

5 Designing Recommendations from Data

Given our descriptive framework connecting recommendation algorithm design to the human decision-maker's counterfactual decisions, an immediate prescriptive question arises: how should we design recommendation algorithms in practice based on observed data only? Above, we showed the value of designing recommendation algorithms that achieve complementarity in a lab experiment where we know the joint distribution of the human's information and the algorithm's information. We now study how our approach can be helpful in the realistic case where we want to design recommendations based on baseline data. To do so, we first outline challenges that arise when we train recommendation algorithms when only unassisted decision data is available. We then propose a maximin approach to address these challenges. Our maximin solution has a particularly intuitive structure: it yields an algorithm that learns where the human decision-maker is likely to make a

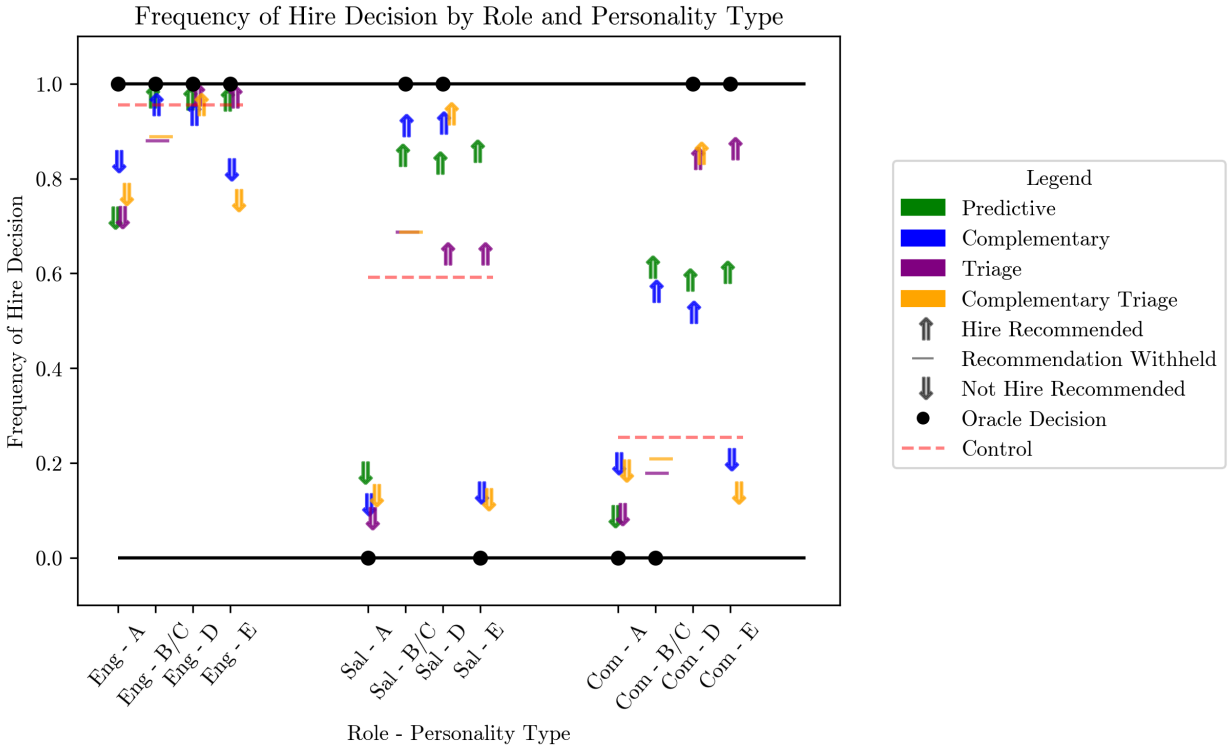


Figure 6: Frequency of experiment subjects taking a hire decision (y-axis) across the different combinations of treatments (color) and profiles (x-axis). Different markers are used to identify whether the subject received a hire recommendation ($\Uparrow$), did not receive a recommendation ($-$), or received a not-hire recommendation ($\Downarrow$). The optimal decision given oracle information is given by the location of the points on the lines $y = 1$ (hire is optimal) and $y = 0$ (not hire is optimal). The dashed lines give the frequency of an applicant being hired by a subject without access to recommendations (*Control* arm).

mistake and then suggests an alternative course of action in those cases. We then consider an alternative strategy that is motivated by the connection to instrumental variables, an intent-to-treat strategy based on randomized recommendations, link the solutions back to the experiment from the previous section, and discuss machine-learning implementations. All strategies are instances of one problem. The designer would like to know the joint distribution of $(Y(1), Y(0), X, D_f(\cdot))$, which under Assumption 2 is $(Y(1), Y(0), X, A_f, C_f)$, and chooses the rule that is best in the worst case over the distributions consistent with what is known,

$$\max_{f \in \mathcal{F}} \min_{\mathbb{P} \in \mathcal{P}} \mathbb{E}[Y(D_f(R))] \quad \text{for } R = f(X), \tag{7}$$

where $\mathcal{P}$ is the set of distributions over $(Y(1), Y(0), X, D_f(\cdot))_{f \in \mathcal{F}}$ that are consistent with the available data and the maintained assumptions. The strategies differ in which assumptions and which data they combine. The assumptions are monotonicity (Assumption 2) and three assumptions introduced below, on the incidence of compliance, on the quality of active decisions, and on invariance across rules. The data conditions are *observed outcomes*, where the joint distribution of $(Y(1), Y(0), X, \underline{D})$ is known, as in the experiment, and a *randomized experiment*, where decisions and outcomes are observed under several randomized rules.

5.1 Challenges in Learning Recommendation Algorithms from Training Data

We consider the task of learning a good recommendation algorithm from baseline training data in which we observe characteristics X, potential outcomes $(Y(1), Y(0))$, as well as unassisted human decisions $\underline{D}$, taken in the absence of algorithmic recommendations (corresponding to the *Control* group in the experiment). We abstract from the task of learning that distribution and assume for now that we know the full joint distribution of

$$(Y(1), Y(0), X, \underline{D}).$$

This is the observed-outcomes data condition. Knowing this distribution, how can we design a good recommendation algorithm?

In general, finding an optimal recommendation algorithm from this baseline data alone is not feasible since relevant parts of the potential-decisions distribution are not identified based on this data alone. Just assuming monotonicity (Assumption 2) does not alleviate the problem. Specifically, we face the following four practical challenges:

1. *No information about compliance.* Since this baseline data does not include a recommendation algorithm, we cannot learn about compliance. Below, we, therefore, make assumptions about the incidence of compliance that express the idea that the decision-maker is more likely to comply in cases where they would have made a mistake, which we justify by relating mistakes

to decision-maker uncertainty.

2. *Recommendations may affect active decisions.* The introduction of a recommendation algorithm may not just affect decisions for compliers, it may also affect active decisions relative to unassisted decisions. Below, we specifically assume that active decisions may improve over unassisted decisions since *not* giving a recommendation may itself provide additional information to the decision-maker.

3. *Compliance and active decisions may vary with the recommendation algorithm.* Even if we can make informed guesses about compliance and active decisions for a specific recommendation algorithm, both may change when a different set of recommendations is given. This phenomenon is related to the concept of *interference* in causal inference: the realization of outcomes (here: decisions) for a specific instance may not only depend on the treatment (here: recommendation) for this instance, but also on the treatment status of other instances. In our formalism, we express this challenge by the explicit dependence of potential decisions $D_f(\cdot)$ (as well as compliance $C = C(f)$ and active decisions $A = A(f)$) on the recommendation algorithm $f \in \mathcal{F}$, such as in Assumption 2 and Proposition 1. Below, we address this challenge by making assumptions that hold across (reasonable) recommendation algorithms $\mathcal{F}$.

4. *Outcomes are selectively observed.* Data on unassisted decisions reveal $Y(1)$ only where $\underline{D} = 1$ and $Y(0)$ only where $\underline{D} = 0$, and $\underline{D}$ depends on the private information U.

These challenges imply that we are not generally able to solve for optimal recommendation algorithms based on baseline data alone. Instead, we have to combine assumptions on compliance, active decisions, and interference. In the spirit of our overall framework, we aim to make relatively weak assumptions. This means there is remaining uncertainty, which we resolve by searching for maximin solutions that make worst-case assumptions about unidentified parts of the distribution of potential decisions.

5.2 Learning from Mistakes with Observed Outcomes

We first consider the observed-outcomes case, in which the designer knows the joint distribution of $(Y(1), Y(0), X, \underline{D})$. Define the *mistake* of the unassisted decision as the gain from reversing it,

$$M = Y(1 - \underline{D}) - Y(\underline{D}),$$

which equals $\tau = Y(1) - Y(0)$ when $\underline{D} = 0$ and $-\tau$ when $\underline{D} = 1$, and write $m_0(x) = \mathbb{E}[M \mid \underline{D} = 0, X = x]$, $m_1(x) = \mathbb{E}[M \mid \underline{D} = 1, X = x]$ and $p_d(x) = \mathbb{P}(\underline{D} = d \mid X = x)$. In the observed-outcomes case these objects are known. A mistake in this sense is a causal effect: the effect of overturning the

human's decision for the instances the human treated in a given way. We impose two assumptions, one on which instances comply and one on active decisions.

Assumption 3 (Incidence of compliance)**.** *For all $f \in \mathcal{F}$, $d \in \{0,1\}$ and x, compliance is non-negatively correlated with mistakes and non-trivial:* $\mathbb{C}\mathrm{ov}(C_f, M \mid \underline{D} = d, X = x) \geq 0$ *and* $\mathbb{P}(C_f = 1 \mid \underline{D} = d, X = x) \geq \varepsilon$ *for some* $\varepsilon > 0$.

The assumption expresses the idea that the decision-maker is more likely to change their decision in response to a recommendation in cases where they would have made a mistake without assistance.

Assumption 4 (Active vs baseline decision quality)**.** *For all recommendation algorithms $f \in \mathcal{F}$,*

$$\mathbb{E}\left[Y(A) - Y(\underline{D}) | C{=}0 \text{ or } R{=}\varnothing\right] \geq 0.$$

Proposition 2 (Learning from mistakes)**.** *Under Assumption 2, Assumption 3 and Assumption 4, and with the joint distribution of $(Y(1), Y(0), X, \underline{D})$ known, the worst-case gain over the unassisted baseline of recommending r at x is $\varepsilon\,\Psi(r,x)$ if $\Psi(r,x) > 0$ and $\Psi(r,x)$ otherwise, where $\Psi(\varnothing, x) = 0$, $\Psi(1,x) = p_0(x)\,m_0(x)$ and $\Psi(0,x) = p_1(x)\,m_1(x)$. The rule*

$$f^{\mathrm{LFM}}(x) \in \operatorname*{arg\,max}_{r \in \{0,\varnothing,1\}} \Psi(r,x)$$

is therefore maximin optimal in (7): it recommends treatment where the instances the human leaves untreated would on average gain from treatment, recommends no treatment where the instances the human treats would on average gain from no treatment, and otherwise sends no recommendation.

The rule coincides with the triage rule (3) (see also Proposition 3), which is no coincidence: Assumption 3 allows compliance to be uncorrelated with mistakes, and the worst case is attained there.

The binary-outcome case In the hiring example of Section 2, M takes the values $c_I = 1 - c$ for a missed good hire, $c_{II} = c$ for a bad hire, and $-c_{II}$ or $-c_I$ when the unassisted decision was right, so that $\mathbb{E}[M \mid \underline{D} = 0, X = x] \geq 0$ is $\mathbb{P}(Y^* = 1 \mid \underline{D} = 0, X = x) \geq c$. The remainder of this subsection restates the result for this case with the mistake replaced by the indicator $\tilde{M}$ of a mistake and with a parameter κ that lets the correlation between compliance and mistakes be weaker or stronger than in Assumption 3.

To determine plausible recommendation algorithms from baseline data about unassisted decisions $\underline{D}$, characteristics X, and latent outcomes Y^*, we now make additional assumptions about how compliance and active decisions relate to unassisted decisions and their performance in the baseline. We then solve for maximin optimal solutions to resolve additional uncertainty about the full distribution of potential decisions.

Recall that the optimal recommendation algorithm under monotonicity (Assumption 2) maximizes

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[Y(R)\,\mathbf{1}\left[C{=}1 \text{ and } R{\neq}\varnothing\right] + Y(A)\,\mathbf{1}\left[C{=}0 \text{ or } R{=}\varnothing\right]\right]$$

under the continued assumptions from Section 3, as in (5). We now address these two components separately. For the second part, we note that we do not directly observe active decisions A, since these may differ from entirely unassisted decisions $\underline{D}$ through additional information provided by the recommendation algorithm. For example, even the (non-)recommendation $R = \varnothing$ may act as a signal that a human decision-maker may leverage together with their private information. To handle such cases, we assume that active decisions cannot be any worse than unassisted decisions (Assumption 4).

As a direct consequence, we can bound the gain from implementing an arbitrary recommendation algorithm f relative to the unassisted baseline case from below by

$$\mathbb{E}\left[Y - Y(\underline{D})\right] \geq \mathbb{E}\left[\left(Y(R) - Y(\underline{D})\right)\mathbf{1}\left[C{=}1 \text{ and } R{\neq}\varnothing\right]\right].$$

Maximizing the worst-case gain then corresponds to maximizing the remaining gain stemming from compliers. To restrict compliance, we consider its relationship to the decision-maker's mistakes in the baseline data. We define

$$\tilde{M} = \mathbf{1}[Y(1{-}\underline{D}){>}Y(\underline{D})] = M_I + M_{II}, \qquad M_I = \mathbf{1}[Y^*{=}1, \underline{D}{=}0], \qquad M_{II} = \mathbf{1}[Y^*{=}0, \underline{D}{=}1],$$

and call $\tilde{M}$ a *(baseline) mistake*, which covers the mutually exclusive cases of a type-I error (M_I) and a type-II error (M_{II}). Were we to assume that, for a specific baseline decision and recommendation, compliance is positively correlated with baseline mistakes,

$$\mathbb{P}(C = 1|\tilde{M} = 1, \underline{D}, R) \geq \mathbb{P}(C = 1|\tilde{M} = 0, \underline{D}, R), \tag{8}$$

then the decision-maker would be weakly more likely to change their decision in response to a recommendation in cases where they would have made a mistake without any assistance. Thus, those cases in which a lot of the same mistakes happen could be improved by suggesting an action that goes in the opposite direction of those mistakes. Specifically, we will later establish that the recommendation rule

$$R = \begin{cases} 0 & \mathbb{E}[\tilde{M}|\underline{D} = 1, X] \geq \frac{c_I}{c_I + c_{II}} \\ 1 & \mathbb{E}[\tilde{M}|\underline{D} = 0, X] \geq \frac{c_{II}}{c_I + c_{II}} \\ \varnothing & \text{else} \end{cases} \tag{9}$$

is maximin optimal (assuming for simplicity that only one of the two conditions is fulfilled at a time). In words, this rule checks how many mistakes were made in baseline data for a given X and

unassisted decision $\underline{D}$. If the probability of a mistake is so high that taking the opposite decisions would improve performance, then (8) will ensure that recommending that alternative plan of action cannot make things any worse.

We now express the idea that compliance is related to mistakes in the baseline data in a more flexible way. Specifically, we use an assumption that expresses how strongly compliance correlates with baseline mistakes. This allows for both expressing assumptions that are weaker than the one in (8) as well as assuming that the relationship between compliance and mistakes is stronger.

Assumption 5 (Incidence of compliance)**.** *For all recommendation algorithms $f \in \mathcal{F}$,*

$$\mathbb{P}(C = 1|\tilde{M} = 1, \underline{D}, R) \geq \kappa\, \mathbb{P}(C = 1|\tilde{M} = 0, \underline{D}, R).$$

This assumption generalizes (8), which is the special case $\kappa = 1$. The assumption is not restrictive at all for $\kappa = 0$, and for $0 < \kappa < 1$ only assumes that compliance is not creating a large number of new mistakes without also correcting some old mistakes. For $\kappa > 1$, it assumes that compliance is more likely if a mistake would have been committed in the baseline and is indicative of the decision-maker using their private information to comply selectively. As κ grows large, more and more mistakes are overridden by compliant decisions, while relatively fewer new mistakes are added. To give this assumption bite in those cases, we also assume that compliance is positive (at least conditional on mistakes):

Assumption 6 (Non-trivial compliance)**.** *For all recommendation algorithms $f \in \mathcal{F}$,*

$$\mathbb{P}(C = 1|\tilde{M} = 1, \underline{D}, R) \geq \varepsilon.$$

Putting all these assumptions together, we can now derive a maximin optimal recommendation algorithm that generalizes the recommendation rule above.

Algorithm 1 (Learning from mistakes)**.** *For a given $\kappa \geq 0$,*

$$f_\kappa^{\text{LFM}}(x) = \begin{cases} 0 & \text{if only } \mathbb{E}[\tilde{M}|\underline{D} = 1, X = x] \geq \frac{c_I}{c_I + \kappa c_{II}} \\ 1 & \text{if only } \mathbb{E}[\tilde{M}|\underline{D} = 0, X = x] \geq \frac{c_{II}}{\kappa c_I + c_{II}} \\ \varnothing & \text{if neither condition holds} \\ \arg\max_{r \in \{1,0\}} \Delta(r, x) & \text{if both conditions hold} \end{cases}$$

where

$$\Delta(r, x) = \begin{cases} \Big((c_I + c_{II}\kappa)\mathbb{E}[\tilde{M}|\underline{D}{=}1, X{=}x] - c_I\Big)\, \mathbb{P}(\underline{D}{=}1|X = x), & r = 0 \\ \Big((\kappa c_I + c_{II})\mathbb{E}[\tilde{M}|\underline{D}{=}0, X{=}x] - c_{II}\Big)\, \mathbb{P}(\underline{D}{=}0|X = x), & r = 1. \end{cases}$$

This algorithm learns from mistakes: it sends a recommendation only if the unassisted decisions lead to so many mistakes that giving the opposite recommendation can only improve decisions. The

structure of the algorithm may look complicated, but is intuitive. The first two rows check whether recommending 0 or 1 are guaranteed to improve outcomes relative to unassisted decisions, respectively. If neither leads to guaranteed improvements in expected outcomes, no recommendation is given; if both lead to improvements, then the one that leads to the larger guaranteed increase is selected.

Theorem 1. *Under Assumption 2, Assumption 4, Assumption 5 and Assumption 6, if Algorithm 1 is in $\mathcal{F}$, then it is maximin optimal in the sense that it solves, for given $\varepsilon \in (0,1]$ and $\kappa \geq 0$,*

$$\max_{f \in \mathcal{F}} \min_{\mathbb{P} \in \mathcal{P}} \mathbb{E}[Y] \text{ for } R = f(X), D = D_f(R), Y = Y(D)$$

where $\mathcal{P}$ are those distributions over $(Y^*, X, \underline{D}, D_f(\cdot))$ that fulfill the assumptions and are consistent with the observed distribution of baseline data $(Y^*, X, \underline{D})$.

We close this part by comparing the resulting algorithm to the alternatives we discuss in Section 3. Specifically, the automated *decision* from (2) that an algorithm could take is to hire whenever $\mathbb{P}(Y^*{=}1|X{=}x) \geq \frac{c_{II}}{c_I + c_{II}}$. While this algorithm ignores human decisions, our proposed approach focuses instead on mistakes that capture the relative performance of human vs machine decisions. The *triage* rule from (3) decides whether to override human decisions based on whether they perform better or worse than automated algorithmic decisions, and is a special case of Algorithm 1.

Proposition 3 (Triage as learning from mistakes)**.** *If $\kappa = 1$, then $f_\kappa^{LFM} = f^{\text{triage}}$.*

The intuition behind this equivalence can be seen by comparing Algorithm 1 to the following representation of the triage rule,

$$f^{\text{triage}}(x) = \arg\max_{r \in \{0,1,\varnothing\}} \begin{cases} -c_I \mathbb{P}(Y^*{=}1|X{=}x), & r = 0 \\ -c_{II} \mathbb{P}(Y^*{=}0|X{=}x), & r = 1 \\ -c_I \mathbb{E}[M_I|X{=}x] - c_{II}\mathbb{E}[M_{II}|X{=}x], & r = \varnothing, \end{cases} \tag{10}$$

where M_I and M_{II} express baseline human type-I and type-II errors as above. Specifically, both formulations show that for a recommendation to be sent, the baseline data must contain a sufficient number of mistakes. If a large number of those mistakes are type-I errors, then both the probability of a mistake given we do not hire and the probability the applicant is good will be high, leading both algorithms to send a hire recommendation. A similar argument holds for type-II errors.

Through Proposition 3, our approach delivers insights into the connection between recommendations and triage:

- If we believe that there is no correlation between compliance and mistakes ($\kappa = 1$), then optimizing for triage yields a maximin optimal recommendation rule within our framework.

- If we believe, however, that compliance is selective in that it is more likely to overturn mistakes than create new ones ($\kappa > 1$), then we obtain a rule that deviates from triage by giving recommendations more often in response to mistakes the unassisted human makes.
- If, on the other hand, we are worried that compliance may not be positively correlated with mistakes ($\kappa < 1$), then we may want to withhold more recommendations.

Overall, these distinctions showcase that the central questions in designing algorithms for complementarity are around how the decision-maker leverages their private information to decide when to comply with a recommendation and how their active decision changes in response to changes in the recommendation algorithm.

5.3 Intent to Treat with Randomized Recommendations

The second strategy replaces assumptions on the incidence of compliance by data on compliance. Suppose the designer has deployed K recommendation rules $f_1, \ldots, f_K \in \mathcal{F}$, one of which may be the null rule $f^{\varnothing}$, assigning each instance to one rule at random, and observes (Y, D, X) together with the rule that was in place. This is the data condition of a randomized experiment, and it is the design of Section 4. The strategy needs the response not to depend on which rule is deployed.

Assumption 7 (Invariance across recommendation rules)**.** *For all $f, f' \in \mathcal{F}$, $(A_f, C_f) = (A_{f'}, C_{f'})$ almost surely.*

Under Assumption 7 we drop the subscript and write (A, C). Under it, the value of each rule at x is identified by the intent-to-treat comparison across arms, and the designer can choose, for each x, the recommendation that performed best.

Proposition 4 (Intent-to-treat strategy)**.** *Suppose Assumption 2 and Assumption 7 hold, and let $V_k(x) = \mathbb{E}[Y \mid X = x, \text{rule } f_k]$ denote the average outcome at x under rule f_k in the experiment. Among all rules that at each x send one of the recommendations $f_1(x), \ldots, f_K(x)$, the rule*

$$f^{\mathrm{ITT}}(x) = f_{k^*(x)}(x), \qquad k^*(x) \in \operatorname*{arg\,max}_{k \in \{1,\ldots,K\}} V_k(x),$$

maximizes $\mathbb{E}[Y(D_f(R))]$.

5.4 Implementation in the Experiment

Above, we have derived maximin algorithms that lead to a guaranteed improvement in expected outcomes under assumptions on compliance and active decisions. We now apply this family of recommendation algorithms to the experiment in Section 4. This application demonstrates both the features and the limitations of the agnostic maximin approach.

	Personality type (x)				
	A	B	C	D	E
$\mathbb{P}(X = x)$	.20	.20	.20	.20	.20
$\mathbb{P}(\underline{D} = 1 \mid X = x)$	0.67	0.68	0.65	0.68	0.67
$\mathbb{P}(Y^* = 1 \mid X = x, \underline{D} = 0)$	0.06	0.52	0.58	1.00	0.53
$\mathbb{P}(Y^* = 0 \mid X = x, \underline{D} = 1)$	0.43	0.07	0.08	0.00	0.36

Table 6: For each personality type x, we report the fraction of applicants, the probability that a subject in the *Control* treatment of our experiment hired an applicant, the probability an applicant was good given they were not hired by a subject in the *Control* treatment (type-I error), and the probability an applicant was good given they were hired by a subject in the *Control* treatment (type-II error).

For our application, we assume that we do *not* have any of the information about private information and its relationship to outcomes available, and instead only rely on information about the joint distribution of decisions, characteristics, and outcomes from the baseline (*Control*) arm. We see this as a conservative reference point that represents the case where we have no or little information about the setting for which we design a recommendation algorithm, making the assumptions from above appropriate.

Per our result above, the main input to finding maximin optimal recommendations using baseline data is the distribution of *mistakes* across characteristics. To calculate the relative improvement of different recommendations, we also require information about unassisted decisions and the distribution of characteristics. These data are represented in Table 6. Together, these numbers completely describe the (empirical) baseline distribution. Here, we take this empirical distribution as given and derive the (sample analog of the) algorithms from Theorem 1 for varying values of κ.

The resulting maximin optimal algorithms are represented in Figure 7 for varying values of κ. Even in the very conservative case of $\kappa \to 0$, the algorithm recommends hiring applicants of personality type D, since any deviation can only be a mistake. In this regime, no other recommendations are given. As κ increases, hiring recommendations are also given for types B, C, and E. Around $\kappa = 1$, this leads to the triage solution from (10). As the minimum correlation between compliance and baseline mistakes increases further, type A is recommended not to be hired (recovering the *Predictive* algorithm), and the recommendation for type E is eventually switched from hiring to not hiring (recovering the *Complementary* algorithm that performs best in the experiment). This connects to the analysis of different algorithms in the experiment in Section 4 and especially in Table 4: if we are willing to assume that compliance is selective (which we interpret here as a high κ), then we expect the *Complementary* algorithm to work well.

Note the approach in this section does not have access to any of the information about complementarity that the algorithms in Section 4 relied on. Learning such additional information and

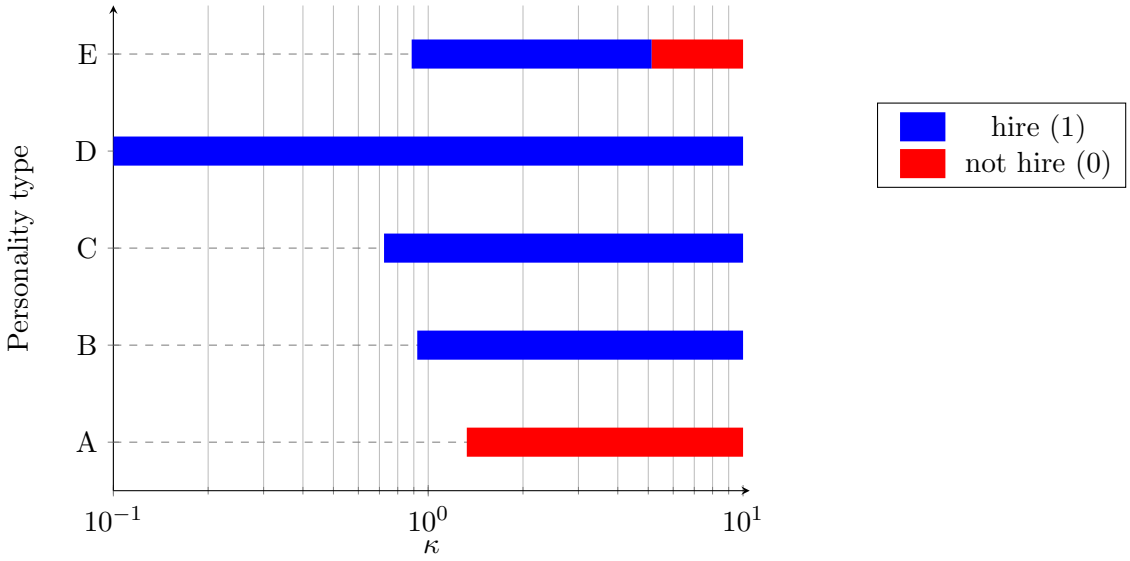


Figure 7: Recommendations provided by Algorithm 1 for varying values of κ based on the *Control* arm of the experiment. Note that the x-axis is on a log scale to highlight the most relevant transitions and because κ is a multiplicative factor that can range from 0 to infinity.

reasoning about compliance and active decisions beyond the arguments above would require access to data about how decisions change with different recommendations, such as those from the treatment arms other than the *Control*.

Overall, this application shows both strengths and limitations of the approach behind Algorithm 1. Even with limited data, our framework can derive sensible recommendation algorithms. At the same time, the approach is inherently limited since it does not rely on any learning about how decisions change in response to recommendations in the given setting and instead is sensitive to specific assumptions. As a result, we obtain both algorithms that improve only somewhat over the baseline (like the *Predictive* algorithm we recover for intermediate κ) or improve strongly (the *Complementary* algorithm for high κ), but could not have told based on baseline data alone which algorithm would have done better. Ultimately, finding better recommendation algorithms will involve adaptive experimentation that learns about the structure of compliance and active decisions. The intent-to-treat strategy of Section 5.3 is the first step of such a program. With the five arms of the experiment as the deployed rules, each personality type received two or three distinct recommendations across arms, and choosing for each type the recommendation of the arm with the highest accuracy on that type gives the rule in Table 7: not hire for type A, no recommendation for B and C, and hire for D and E, which is the *Triage* algorithm. Its validity rests on Assumption 7.

Personality type	Recommendations across arms	Intent-to-treat rule
A	∅, 0	0
B	1, ∅	∅
C	1, ∅	∅
D	1, ∅	1
E	1, ∅, 0	1

Table 7: Intent-to-treat rule from the experiment. For each personality type, the recommendations sent by the five arms (Table 3) and the recommendation of the arm with the highest accuracy on that type.

5.5 Machine-Learning Implementation

We close this section by discussing the practical implementation of recommendation algorithms using machine learning. We assumed above knowledge of the joint distribution of unassisted decisions, characteristics, and outcomes. In practice, these relationships would have to be learned from data. Fortunately, the algorithm above is instructive about how this learning could be implemented effectively. Specifically, the main ingredients to constructing Algorithm 1 are the probability of mistakes conditional on characteristics and unassisted decisions as well as the baseline decisions. These correspond to the solution of straightforward classification problems that estimate

$$m_0(x) = \mathbb{E}[\tilde{M}|\underline{D}{=}0, X{=}x], \qquad m_1(x) = \mathbb{E}[\tilde{M}|\underline{D}{=}1, X{=}x], \qquad h(x) = \mathbb{P}(\underline{D} = 1|X{=}x).$$

Based on estimates $\hat{m}_0, \hat{m}_1, \hat{h}$, a plug-in rule is

$$\hat{f}(x) = \underset{r\in\{0,1,\varnothing\}}{\arg\max} \begin{cases} ((c_I + \kappa c_{II})\hat{m}_1(x) - c_I)\hat{h}(x), & r = 0 \\ ((\kappa c_I + c_{II})\hat{m}_0(x) - c_{II})(1 - \hat{h}(x)), & r = 1 \\ 0, & r = \varnothing. \end{cases}$$

An alternative is to estimate the rule directly via cost-weighted classification by solving the sample analog of

$$\max_{f\in\mathcal{F}} \mathbb{E}\left[\mathbf{1}[f(X){=}1]((\kappa c_I + c_{II})M_I - c_{II}\mathbf{1}[\underline{D}{=}0]) + \mathbf{1}[f(X){=}0]((c_I + \kappa c_{II})M_{II} - c_I\mathbf{1}[\underline{D}{=}1])\right]$$

where M_I and M_{II} are type-I and type-II errors, respectively.

6 Conclusion

This article formalizes the analysis and optimal design of recommendation algorithms as a causal design problem with endogenous response, which employs the potential-outcomes framework from

causal inference. We draw similarities between the impact that recommendation algorithms have on human decisions in algorithmic design and the effect that instrumental variables have on individual treatment choices in inference problems. We use this connection to decompose the objective function of a recommendation algorithm into algorithmic performance and human response. Making additional assumptions on reasonable human responses to recommendations, our approach allows us to express recommendation-assisted decisions in terms of compliance and active decisions. We can then reason about optimal recommendation algorithms in terms of assumptions on each component individually. We demonstrate the utility of our framework in a controlled online experiment and a maximin optimization application, both of which reinforce that decision-maker responses should be considered when designing recommendation algorithms. Beyond learning from mistakes, randomized recommendations yield an intent-to-treat strategy, with an explicit list of the assumptions and data it needs. Specifically, we document in the experiment that subjects perform better using recommendations that are designed to provide complementary information, relative to algorithms that are instead designed to automate decisions.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. (Cited on page 4.)

Angelova, Victoria, Will S. Dobbie, and Crystal Yang (2023). Algorithmic Recommendations and Human Discretion. (Cited on page 4.)

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455. (Cited on pages 5, 10, and 12.)

Athey, Susan and Stefan Wager (2021). Policy Learning With Observational Data. *Econometrica*, 89(1):133–161. (Cited on page 5.)

Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans (2020). The Allocation of Decision Authority to Human and Artificial Intelligence. *AEA Papers and Proceedings*, 110:80–84. (Cited on pages 5 and 15.)

Baek, Jackie, Justin J. Boutilier, Vivek F. Farias, Jonas Oddur Jonasson, and Erez Yoeli (2023). Policy Optimization for Personalized Interventions in Behavioral Health. arXiv:2303.12206 [cs]. (Cited on page 5.)

Balakrishnan, Maya, Kris Ferreira, and Jordan Tong (2022). Improving Human-Algorithm Collaboration: Causes and Mitigation of Over- and Under-Adherence. (Cited on page 4.)

Bansal, Gagan, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:2–11. (Cited on pages 4 and 26.)

Bastani, Hamsa, Osbert Bastani, and Wichinpong Park Sinchaisri (2022). Improving Human Decision-Making with Machine Learning. arXiv:2108.08454 [cs]. (Cited on page 5.)

Ben-Michael, Eli, D. James Greiner, Melody Huang, Kosuke Imai, Zhichao Jiang, and Sooahn Shin (2025). Does AI help humans make better decisions? A statistical evaluation framework for experimental and observational studies. *Proceedings of the National Academy of Sciences*, 122(38):e2505106122. (Cited on page 5.)

Bergemann, Dirk and Stephen Morris (2019). Information Design: A Unified Perspective. *Journal of Economic Literature*, 57(1):44–95. (Cited on page 8.)

Cadigan, Timothy and Christopher Lowenkamp (2011). Implementing risk assessment in the federal pretrial services system. *Federal Probation Journal*, 75(2). (Cited on page 2.)

Caro, Felipe and Anna Sáez de Tejada Cuenca (2023). Believing in Analytics: Managers' Adherence to Price Recommendations from a DSS. *Manufacturing & Service Operations Management*, page msom.2022.1166. (Cited on page 5.)

Cui, Yifan and Eric Tchetgen Tchetgen (2021). A Semiparametric Instrumental Variable Approach to Optimal Treatment Regimes Under Endogeneity. *Journal of the American Statistical Association*, 116(533):162–173. (Cited on page 5.)

Donahue, Kate, Alexandra Chouldechova, and Krishnaram Kenthapadi (2022). Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. arXiv:2202.08821 [cs]. (Cited on page 4.)

Fogliato, Riccardo, Maria De-Arteaga, and Alexandra Chouldechova (2022). A Case for Humans-in-the-Loop: Decisions in the Presence of Misestimated Algorithmic Scores. (Cited on page 4.)

Green, Ben and Yiling Chen (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):50:1–50:24. (Cited on page 4.)

Heckman, James J. and Edward Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738. (Cited on page 5.)

Hemmer, Patrick, Max Schemmer, Michael Vössing, and Niklas Kühl (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS 2021 Proceedings*. (Cited on page 2.)

Ibrahim, Rouba, Song-Hee Kim, and Jordan Tong (2021). Eliciting Human Judgment for Prediction Algorithms. *Management Science*, 67(4):2314–2325. (Cited on page 5.)

Imai, Kosuke, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin (2021). Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment. arXiv:2012.02845 [cs, stat]. (Cited on pages 4 and 5.)

Imbens, Guido W. and Joshua D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467. (Cited on pages 3, 5, and 9.)

Kamenica, Emir and Matthew Gentzkow (2011). Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615. (Cited on page 8.)

Kitagawa, Toru and Aleksey Tetenov (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616. (Cited on page 5.)

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*. (Cited on page 5.)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495. (Cited on page 7.)

Lai, Vivian, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan (2021). Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. arXiv:2112.11471 [cs]. (Cited on page 4.)

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 275–284, New York, NY, USA. Association for Computing Machinery. (Cited on page 5.)

Lee, Sokbae and Bernard Salanié (2018). Identifying effects of multivalued treatments. *Econometrica*, 86(6):1939–1963. (Cited on page 12.)

Li, Lan, Tina Lassiter, Joohee Oh, and Min Kyung Lee (2021). Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 166–176, New York, NY, USA. Association for Computing Machinery. (Cited on page 12.)

Manski, Charles F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246. (Cited on page 5.)

Maron, Roman C., Jochen S. Utikal, Achim Hekler, Axel Hauschild, Elke Sattler, Wiebke Sondermann, Sebastian Haferkamp, Bastian Schilling, Markus V. Heppt, Philipp Jansen, Markus Reinholz, Cindy Franklin, Laurenz Schmitt, Daniela Hartmann, Eva Krieghoff-Henning, Max Schmitt, Michael Weichenthal, Christof von Kalle, Stefan Fröhling, and Titus J. Brinker (2020). Artificial Intelligence and Its Effect on Dermatologists' Accuracy in Dermoscopic Melanoma Image Classification: Web-Based Survey Study. *Journal of Medical Internet Research*, 22(9):e18091. (Cited on page 4.)

McLaughlin, Bryce and Jann Spiess (2022). Algorithmic Assistance with Recommendation-Dependent Preferences. arXiv:2208.07626 [cs, econ, q-fin]. (Cited on pages 5 and 9.)

Mills, Chris and Marie-Pascale Grimon (2022). The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial. (Cited on page 4.)

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky (2018). Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. *Econometrica*, 86(5):1589–1619. (Cited on page 5.)

Mogstad, Magne and Alexander Torgovitsky (2018). Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics*, 10(1):577–613. (Cited on page 5.)

Noti, Gali and Yiling Chen (2023). Learning When to Advise Human Decision Makers. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3038–3048, Macau, SAR China. International Joint Conferences on Artificial Intelligence Organization. (Cited on pages 5 and 9.)

Orfanoudaki, Agni, Soroush Saghafian, Karen Song, Harini A. Chakkera, and Curtiss B. Cook (2022). Algorithm, Human, or the Centaur: How to Enhance Clinical Care? (Cited on page 4.)

Pu, Hongming and Bo Zhang (2021). Estimating Optimal Treatment Rules with an Instrumental Variable: A Partial Identification Learning Approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):318–345. (Cited on page 5.)

Qiu, Hongxiang, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C. Kessler, and Alex Luedtke (2021). Optimal Individualized Decision Rules Using Instrumental Variable Methods. *Journal of the American Statistical Association*, 116(533):174–191. (Cited on page 5.)

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 469–481, New York, NY, USA. Association for Computing Machinery. (Cited on pages 2 and 8.)

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan (2019). The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. arXiv:1903.12220 [cs]. (Cited on pages 14 and 15.)

Rambachan, Ashesh (2024). Identifying Prediction Mistakes in Observational Data. *The Quarterly Journal of Economics*, 139(3):1665–1711. (Cited on page 5.)

Rubin, Donald (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701. (Cited on page 9.)

Snyder, Clare, Samantha Keppler, and Stephen Leider (2023). Algorithm Reliance Under Pressure: The Effect of Customer Load on Service Workers. (Cited on page 5.)

Stevenson, Megan T. and Jennifer L. Doleac (2022). Algorithmic Risk Assessment in the Hands of Humans. (Cited on page 4.)

Sun, Jiankun, Dennis J. Zhang, Haoyuan Hu, and Jan A. Van Mieghem (2022). Predicting Human Discretion to Adjust Algorithmic Prescription: A Large-Scale Field Experiment in Warehouse Operations. *Management Science*, 68(2):846–865. (Cited on page 5.)

Sutton, Reed T., David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):1–10. (Cited on page 2.)

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303. (Cited on page 2.)

Vytlacil, Edward (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, 70(1):331–341. (Cited on page 5.)

Wu, Yifan, Ziyang Guo, Michails Mamakos, Jason Hartline, and Jessica Hullman (2023). The Rational Agent Benchmark for Data Visualization. arXiv:2304.03432 [cs]. (Cited on page 24.)

Appendix

A Proofs

The proofs of Theorem 1 and Proposition 3 are written in the loss notation of Example 1, with $\ell(Y^*, D) = c_I \mathbf{1}[Y^*{=}1, D{=}0] + c_{II}\mathbf{1}[Y^*{=}0, D{=}1]$, and write Y for the ability Y^*. Since $Y(D) = (Y^* - c)D$ with $c_I = 1 - c$ and $c_{II} = c$, we have $\ell(Y^*, D) = c_I Y^* - Y(D)$, so that minimizing expected loss is the same as maximizing the expected outcome, and the mistake indicator $\tilde{M}$ is the one defined in Section 5.2.

Proof of Proposition 1. To show this result we will build a bijection between the potential outcomes $(D_{f,i}(0), D_{f,i}(\varnothing), D_{f,i}(1))$ which exist under Assumption 2 and realizations of (A_i, C_i):

$$\begin{aligned} (0,0,0) &\longleftrightarrow (0,0) \\ (0,0,1) &\longleftrightarrow (0,1) \\ (0,1,1) &\longleftrightarrow (1,1) \\ (1,1,1) &\longleftrightarrow (1,0) \end{aligned}$$

□

Proof of Proposition 2. Fix $f \in \mathcal{F}$ and write $R = f(X)$. By Proposition 1,

$$Y(D) - Y(\underline{D}) = C_f \left(Y(R) - Y(\underline{D})\right) \mathbf{1}[R \neq \varnothing] + \left(Y(A_f) - Y(\underline{D})\right) \mathbf{1}[C_f = 0 \text{ or } R = \varnothing].$$

The expectation of the second term is non-negative by Assumption 4, and it equals zero when $A_f = \underline{D}$, which is admissible. For the first term, $Y(1) - Y(\underline{D}) = (1 - \underline{D})\, M$ and $Y(0) - Y(\underline{D}) = \underline{D}\, M$, so that

$$\begin{aligned} \mathbb{E}\left[C_f \left(Y(R) - Y(\underline{D})\right) \mathbf{1}[R \neq \varnothing]\right] = \sum_x \mathbb{P}(X = x) \Big(& \mathbf{1}[f(x) = 1]\, p_0(x)\, \mathbb{E}[C_f M \mid \underline{D} = 0, X = x] \\ & + \mathbf{1}[f(x) = 0]\, p_1(x)\, \mathbb{E}[C_f M \mid \underline{D} = 1, X = x] \Big). \end{aligned}$$

For each d and x, the covariance identity gives

$$\mathbb{E}[C_f M \mid \underline{D} = d, X = x] = \mathbb{C}\mathrm{ov}(C_f, M \mid \underline{D} = d, X = x) + \mathbb{P}(C_f = 1 \mid \underline{D} = d, X = x)\, m_d(x),$$

where the covariance is non-negative and the probability lies in $[\varepsilon, 1]$ by Assumption 3. If $m_d(x) > 0$, the right-hand side is at least $\varepsilon\, m_d(x)$, with equality when C_f is independent of M given $(\underline{D} = d, X = x)$ and $\mathbb{P}(C_f = 1 \mid \underline{D} = d, X = x) = \varepsilon$. If $m_d(x) \leq 0$, it is at least $m_d(x)$, with equality when $C_f = 1$ on $\{\underline{D} = d, X = x\}$. Both choices satisfy Assumption 3, and they can be made separately for each (d, x) and each f and combined with $A_f = \underline{D}$; by Proposition 1 every such specification of

(A_f, C_f) is consistent with Assumption 2, and it leaves the known distribution of $(Y(1), Y(0), X, \underline{D})$ unchanged. Writing $G(r,x) = \varepsilon\Psi(r,x)$ if $\Psi(r,x) > 0$ and $G(r,x) = \Psi(r,x)$ otherwise, we obtain

$$\min_{\mathbb{P}\in\mathcal{P}} \mathbb{E}[Y(D_f(R)) - Y(\underline{D})] = \sum_x \mathbb{P}(X=x)\, G(f(x),x),$$

which is the first claim. The maximin problem (7) therefore reduces to maximizing $G(f(x),x)$ separately for each x. Since $G(r,x)$ and $\Psi(r,x)$ have the same sign for every r, both vanish at $r = \varnothing$, and $G = \varepsilon\Psi$ wherever $\Psi > 0$, the two functions have the same maximizers in r, which proves the second claim. □

Proof of Theorem 1. The optimal recommendation algorithm minimizes

$$\begin{aligned}
&\mathbb{E}\left[\ell(Y, D_f(R)))\right] - \mathbb{E}\left[\ell(Y,\underline{D}))\right] \\
&= \mathbb{E}\left[(\ell(Y,R) - \ell(Y,\underline{D}))\,\mathbf{1}\left(C{=}1 \text{ and } R{\neq}\varnothing\right) + (\ell(Y,A) - \ell(Y,\underline{D}))\,\mathbf{1}\left(C{=}0 \text{ or } R{=}\varnothing\right)\right] \\
&\leq \mathbb{E}\left[(\ell(Y,R) - \ell(Y,\underline{D}))\,\mathbf{1}\left(C{=}1 \text{ and } R{\neq}\varnothing\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(\ell(Y,R) - \ell(Y,\underline{D}))\mathbf{1}\left(C{=}1\right)|R,\underline{D}\right]\mathbf{1}\left(R{\notin}\{\varnothing,\underline{D}\}\right)\right]
\end{aligned}$$

where, for $r \neq d$ and $\bar{p}(r,d) = \mathbb{P}(C{=}1|\tilde{M}{=}0, R{=}r, \underline{D}{=}d)$,

$$\begin{aligned}
&\mathbb{E}\left[(\ell(Y,R) - \ell(Y,\underline{D}))\mathbf{1}\left(C{=}1\right)|R{=}r,\underline{D}{=}d\right] \\
&= \begin{cases}
\overbrace{(-c_{II})}^{\text{gain from mistake corrected}}\mathbb{P}(C{=}1|\tilde{M}{=}1,R{=}r,\underline{D}{=}d)\mathbb{P}(\tilde{M}{=}1|R{=}r,\underline{D}{=}d) \\
+c_I\mathbb{P}(C{=}1|\tilde{M}{=}0,R{=}r,\underline{D}{=}d)\mathbb{P}(\tilde{M}{=}0|R{=}r,\underline{D}{=}d), & d=1, r=0 \\
(-c_I)\mathbb{P}(C{=}1|\tilde{M}{=}1,R{=}r,\underline{D}{=}d)\mathbb{P}(\tilde{M}{=}1|R{=}r,\underline{D}{=}d) \\
+\underbrace{c_{II}}_{\text{loss from new mistake}}\mathbb{P}(C{=}1|\tilde{M}{=}0,R{=}r,\underline{D}{=}d)\mathbb{P}(\tilde{M}{=}0|R{=}r,\underline{D}{=}d), & d=0, r=1
\end{cases} \\
&\leq \bar{p}(r,d)\begin{cases}
-c_{II}\kappa\mathbb{E}[\tilde{M}|R{=}r,\underline{D}{=}d] + c_I(1{-}\mathbb{E}[\tilde{M}|R{=}r,\underline{D}{=}d]) & d=1,r=0 \\
-c_I\kappa\mathbb{E}[\tilde{M}|R{=}r,\underline{D}{=}d] + c_{II}(1{-}\mathbb{E}[\tilde{M}|R{=}r,\underline{D}{=}d]) & d=0,r=1
\end{cases} \\
&= \bar{p}(r,d)\begin{cases}
c_I - (c_I + \kappa c_{II})\mathbb{E}[\tilde{M}|R{=}r,\underline{D}{=}d] & d=1,r=0 \\
c_{II} - (\kappa c_I + c_{II})\mathbb{E}[\tilde{M}|R{=}r,\underline{D}{=}d] & d=0,r=1
\end{cases}
\end{aligned}$$

where the lower bound can be attained by assuming that the compliance bound from Assumption 5

is binding. Writing $m_d(x) = \mathbb{E}[\tilde{M}|X{=}x, \underline{D}{=}d]$, we therefore obtain the tight upper bound

$$\begin{aligned}
&\mathbb{E}\left[\ell(Y, D_f(R)))\right] - \mathbb{E}\left[\ell(Y, \underline{D}))\right] \\
&\leq \mathbb{E}\big[\bar{p}(0,1)\mathbf{1}(\underline{D}{=}1, R{=}0)(c_I - (c_I + \kappa c_{II})m_1(X)) \\
&\qquad + \bar{p}(1,0)\mathbf{1}(\underline{D}{=}0, R{=}1)(c_{II} - (\kappa c_I + c_{II})m_0(X))\big] \\
&= \mathbb{E}\big[\bar{p}(0,1)\mathbf{1}(f(X){=}0)(c_I - (c_I + \kappa c_{II})m_1(X))\mathbb{P}(\underline{D}{=}1|X) \\
&\qquad + \bar{p}(1,0)\mathbf{1}(f(X){=}1)(c_{II} - (\kappa c_I + c_{II})m_0(X))\mathbb{P}(\underline{D}{=}0|X)\big].
\end{aligned}$$

To minimize this upper bound, we would want to assign only those instances x to the 0 recommendation for which $c_I - (c_I + \kappa c_{II})m_1(X)$ is negative and only those instances x to the 1 recommendation for which $c_{II} - (\kappa c_I + c_{II})m_0(X)$ is negative (since we can otherwise assign to $\varnothing$). As a result, all components are non-positive, and the bound from Assumption 6 is binding in the worst case, yielding the (still tight) upper bound (for the non-trivial case of $\kappa > 0$)

$$\begin{aligned}
&\mathbb{E}\left[\ell(Y, D_f(R)))\right] - \mathbb{E}\left[\ell(Y, \underline{D}))\right] \\
&\leq \varepsilon/\kappa\mathbb{E}\big[\mathbf{1}(f(X){=}0)(c_I - (c_I + \kappa c_{II})m_1(X))\mathbb{P}(\underline{D}{=}1|X) \\
&\qquad + \mathbf{1}(f(X){=}1)(c_{II} - (\kappa c_I + c_{II})m_0(X))\mathbb{P}(\underline{D}{=}0|X)\big].
\end{aligned}$$

We now write $h(x) = \mathbb{P}(\underline{D}{=}1|X{=}x)$ For instances x for which both $f(x) \in \{1, 0\}$ yield a reduction in loss, an optimal choice is the maximizer of the gain

$$\Delta(r, x) = \begin{cases} (c_I + \kappa c_{II})m_1(X)h(x) - c_I h(x), & r = 0 \\ (\kappa c_I + c_{II})m_0(X)(1 - h(x)) - c_{II}(1 - h(x)), & r = 1, \end{cases}$$

which together yields Algorithm 1. □

Proof of Proposition 3. To show this result we will show Equation 10 recovers Algorithm 1 for $\kappa = 1$. Do to so we will recover the first three cases of Algorithm 1 by comparing the terms in Equation 10 for $r = 0$ and $r = 1$ to $r = \varnothing$, then recover the last condition by comparing $r = 0$ to $r = 1$.

Here we rewrite the condition for $r = 0$ to be preferred to $r = \varnothing$ as,

$$\begin{aligned}
c_I \mathbb{P}(Y = 1|X = x) &\leq c_I \mathbb{E}\left[M_I|X = x\right] + c_{II}\mathbb{E}\left[M_{II}|X = x\right] \\
c_I \mathbb{P}(Y = 1|X = x) &\leq c_I \mathbb{P}(Y = 1, \underline{D} = 0|X = x) + c_{II}\mathbb{P}(Y = 0, \underline{D} = 1|X = x) \\
c_I \mathbb{P}(Y = 1, \underline{D} = 1|X = x) &\leq c_{II}\left[\mathbb{P}(\underline{D} = 1|X = x) - \mathbb{P}(Y = 1, \underline{D} = 1|X = x)\right] \\
\mathbb{P}(Y = 1|\underline{D} = 1, X = x) &\leq \frac{c_{II}}{c_I + c_{II}} \\
1 - \mathbb{P}(Y = 0|\underline{D} = 1, X = x) &\leq \frac{c_{II}}{c_I + c_{II}} \\
\mathbb{P}(Y = 0|\underline{D} = 1, X = x) &\geq \frac{c_I}{c_I + c_{II}} \\
\mathbb{E}\left[\tilde{M}|\underline{D} = 1, X = x\right] &\geq \frac{c_I}{c_I + c_{II}},
\end{aligned}$$

recovering the first case of Algorithm 1 for $\kappa = 1$. Similarly we can rewrite the condition for $r = 1$ to be preferred to $r = \varnothing$ as,

$$\begin{aligned}
c_{II} \mathbb{P}(Y = 0|X = x) &\leq c_I \mathbb{E}\left[M_I|X = x\right] + c_{II}\mathbb{E}\left[M_{II}|X = x\right] \\
c_{II} \mathbb{P}(Y = 0|X = x) &\leq c_I \mathbb{P}(Y = 1, \underline{D} = 0|X = x) + c_{II}\mathbb{P}(Y = 0, \underline{D} = 1|X = x) \\
c_{II} \mathbb{P}(Y = 0, \underline{D} = 0|X = x) &\leq c_I\left[\mathbb{P}(\underline{D} = 0|X = x) - \mathbb{P}(Y = 0, \underline{D} = 0|X = x)\right] \\
\mathbb{P}(Y = 0|\underline{D} = 0, X = x) &\leq \frac{c_I}{c_I + c_{II}} \\
1 - \mathbb{P}(Y = 1|\underline{D} = 0, X = x) &\leq \frac{c_I}{c_I + c_{II}} \\
\mathbb{P}(Y = 1|\underline{D} = 0, X = x) &\geq \frac{c_{II}}{c_I + c_{II}} \\
\mathbb{E}\left[\tilde{M}|\underline{D} = 0, X = x\right] &\geq \frac{c_{II}}{c_I + c_{II}},
\end{aligned}$$

which recovers the second and third cases of Algorithm 1 for $\kappa = 1$. Now we must consider when both $r = 0$ and $r = 1$ are preferred to $r = \varnothing$ and recover the form of $\Delta(r, x)$ when $\kappa = 1$. To do so we rewrite,

$$\begin{aligned}
c_I \mathbb{P}(Y = 1|X = x) &= c_I\left[\mathbb{P}(Y = 1, \underline{D} = 0|X = x) + \mathbb{P}(Y = 1, \underline{D} = 1|X = x)\right] \\
&= c_I\left[\mathbb{E}\left[M_I|X = x\right] + \mathbb{P}(\underline{D} = 1|X = x) - \mathbb{E}\left[M_{II}|X = x\right]\right]
\end{aligned}$$

and

$$\begin{aligned}
c_{II} \mathbb{P}(Y = 0|X = x) &= c_{II}\left[\mathbb{P}(Y = 0, \underline{D} = 1|X = x) + \mathbb{P}(Y = 0, \underline{D} = 0|X = x)\right] \\
&= c_{II}\left[\mathbb{E}\left[M_{II}|X = x\right] + \mathbb{P}(\underline{D} = 0|X = x) - \mathbb{E}\left[M_I|X = x\right]\right]
\end{aligned}$$

and subtract $c_I\mathbb{E}[M_I|X = x] + c_{II}\mathbb{E}[M_{II}|X = x]$ (the value of $r = \varnothing$), from each to recover and

equivalent representation of $\Delta(r, x)$. □

Proof of Proposition 4. Under Assumption 7, write (A, C) for the common value of (A_f, C_f) over $f \in \mathcal{F}$. By Proposition 1, for every rule g the realized outcome is $Y(D_g(g(X))) = h(g(X))$, where

$$h(r) = Y(r)\,\mathbf{1}[r \neq \varnothing,\, C = 1] + Y(A)\,\mathbf{1}[r = \varnothing \text{ or } C = 0]$$

is a function of r and of $(Y(1), Y(0), A, C)$ alone, and does not depend on the rule. Hence, with $W(r, x) = \mathbb{E}[h(r) \mid X = x]$,

$$\mathbb{E}[Y(D_g(g(X)))] = \mathbb{E}\big[W(g(X), X)\big],$$

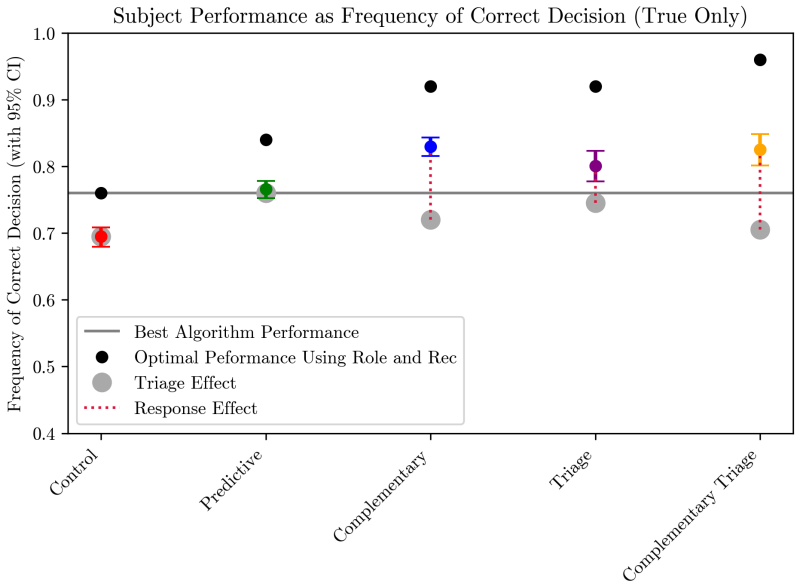
and a rule that at each x chooses $g(x)$ from the set $\{f_1(x), \ldots, f_K(x)\}$ to maximize $W(g(x), x)$ maximizes the objective among all rules with this choice set. It remains to identify $W(f_k(x), x)$. In the experiment, the rule f_k is assigned independently of the instance, so the conditional distribution of $(Y(1), Y(0), A, C)$ given $X = x$ is the same in every arm, and the realized outcome in arm k is $h(f_k(X))$ by Assumption 1. Therefore $V_k(x) = \mathbb{E}[Y \mid X = x, \text{rule } f_k] = \mathbb{E}[h(f_k(x)) \mid X = x] = W(f_k(x), x)$, and $k^*(x) \in \arg\max_k V_k(x)$ selects a recommendation $f_{k^*(x)}(x)$ that maximizes $W(\cdot, x)$ over the deployed recommendations at x. The proof uses Assumption 2 only through the representation of Proposition 1; invariance of the potential decisions $D_f(\cdot)$ across rules would suffice in its place. □

B Robustness Checks

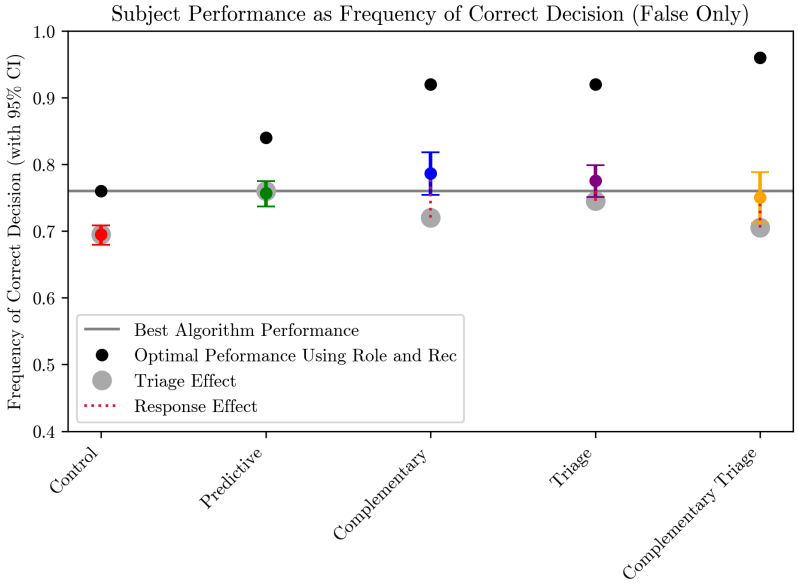
Here we reproduce the results of Figure 5 and Figure 6 for the sub-populations who answered all of their recommendation comprehension questions correctly (585 subjects) and those who answered at least one question wrong (376 subjects). Subjects who answered at least one recommendation comprehension question incorrectly exhibited noisier decisions on average than those who answered all questions correctly. We find that algorithms generated using our framework exhibited higher levels of complementarity for the sub-population who correctly answered all comprehension questions while the sub-population who answered at least one comprehension question wrong failed to realize complementarity.

Figure 8 reports each sub-populations average performance by treatment. Both sub-populations performed equally well using the *Predictive* algorithm; however, only the sub-population which answered all of our recommendation comprehension questions correctly was able to further improve their performance using the *Triage*, *Complementary*, and *Complementary Triage* algorithms. Performing well with these algorithms required subjects to effectively parse when the recommendation was accurate or erroneous given their private information, thus we expect (and observe) that understanding how the algorithm generates recommendations is vital for its effective usage.

Figure 9 reports how each sub-populations propensity to hire applications varied across the



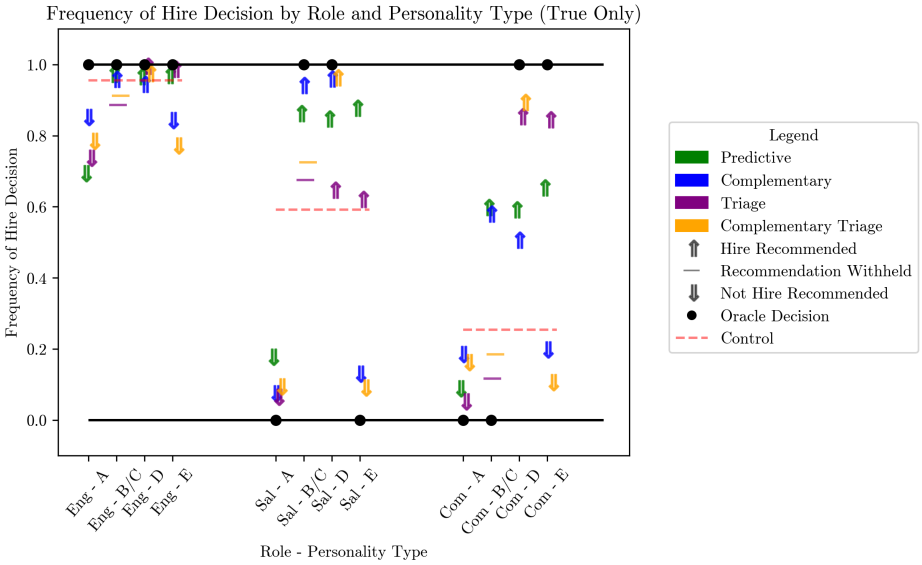
(a) Sub-population with complete comprehension



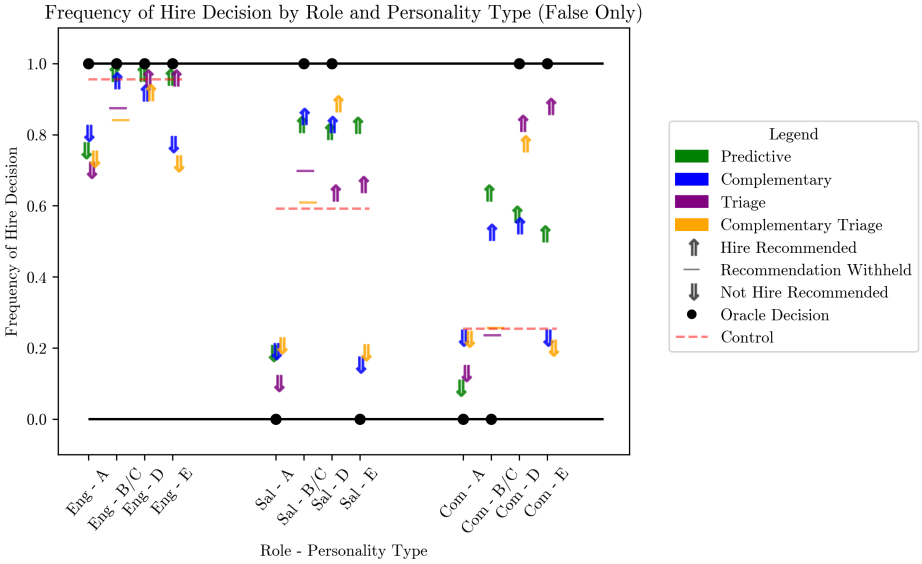
(b) Sub-population with comprehension errors

Figure 8: Average subject performance (through fractions of optimal hiring decisions) across treatments. 95% bootstrap confidence intervals are clustered at the subject level. In addition, the triage performance (where compliance is perfect and not sending a recommendation uses the decision of *Control*) is given for each treatment by a gray dot, while the Response Effect from Section 3.5 is shown by the vertical red dashed lines. The optimal algorithm performance is shown by the gray horizontal line. The optimal performance achievable by a rational agent with the subject's knowledge (role and algorithmic recommendation) is given for each treatment arm as a black dot.

profiles they viewed. Generally subjects who answered at least one recommendation comprehension question wrong exhibited noisier decisions. This difference is particularly noticeable on Sales profiles for subjects in the *Complementary* and *Complementary Triage* treatments. Other than this excess noise, subjects who answered at least one recommendation comprehension question wrong responded to recommendations in a similar way to subjects who answered all recommendation comprehension questions correctly.



(a) Sub-population with complete comprehension



(b) Sub-population with comprehension errors

Figure 9: Frequency of experiment subjects taking a hire decision (y-axis) across the different combinations of treatments (color) and profiles (x-axis). Different markers are used to identify whether the subject received a hire recommendation ($\Uparrow$), did not receive a recommendation ($-$), or received a not-hire recommendation ($\Downarrow$). The optimal decision given oracle information is given by the location of the points on the lines $y = 1$ (hire is optimal) and $y = 0$ (not hire is optimal). The dashed lines give the frequency of an applicant being hired by a subject without access to recommendations (*Control* arm). The legend for these graphs is the same as the one in Figure 6 in the main text.