The Duet of Representations and How Explanations Exacerbate It (Short Paper)

CHARLES WAN, Rotterdam School of Management, Erasmus University, The Netherlands
RODRIGO BELO, Nova School of Business and Economics, Universidade NOVA de Lisboa, Portugal
LEID ZEJNILOVIĆ, Nova School of Business and Economics, Universidade NOVA de Lisboa, Portugal
SUSANA LAVADO, Nova School of Business and Economics, Universidade NOVA de Lisboa, Portugal

An algorithm effects a causal representation of relations between features and labels. This representation might conflict with the human's prior belief. Explanations can direct the human's attention to the conflicting feature and away from other relevant features. This leads to causal overattribution and may adversely affect the human's information processing. We present empirical evidences that the quality of the human's decision-making is worse when a feature on which the human holds a conflicting prior belief is displayed as part of the explanation.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Empirical studies in HCI**; **HCI theory, concepts and models**.

Additional Key Words and Phrases: human-AI interaction, communication, representations, priors, biases, explanations, epistemic standpoint, salience, information processing

1 INTRODUCTION

Artificial intelligence is increasingly embedded in everyday business and consumer decision-making. For an array of reasons — technical, psychological, organizational, legal, and ethical — these decision-making systems are seldom fully automated. They require human input or, as components of larger socio-technical systems, interface to a significant degree with other components that are predominantly human-driven. Understanding how humans interact with algorithms *epistemically*, therefore, is of crucial importance whether the considerations are primarily economic and managerial or societal and ethical.

We argue that one principal way algorithms communicate with humans is via causal representations. The algorithm effects a causal representation that relates features and labels [DeLanda 2021]. This is conveyed to the human through the algorithm's observable output — typically predictions. For example, upon observing the algorithm's output (predictions), the human might attribute to the algorithm a causal representation with the following semantics: longer job tenure leads to a lower risk of loan default.

Informed by history, humans form prior beliefs with respect to predictive tasks. For example, in predicting loan default humans might associate lower income with a higher risk of loan default. This representation may conflict with the causal representation attributed to the algorithm. We call the conflict the duet of representations. Because humans value simplicity, both their prior beliefs and the causal representations they attribute to algorithms are likely to be linear and sparse [Lombrozo 2007].

One way to facilitate communication between agents is through the provision of explanations. Explanatory methods can be viewed as tools that render the algorithm's causal representations human-interpretable. They extract simple representations via some model additive to the original algorithm. For example, LIME [Ribeiro et al. 2016] use locally linear models to extract linear representations of relations between features and labels. SHAP [Lundberg and Lee 2017] use cooperative games with features as players and extract representations in the form of sets of important features. Explanations direct human *attention* to sparse and cogent representations with clear semantics. They therefore increase

the salience of any conflict with human priors. We present empirical evidences from a field experiment that explanations exacerbate the duet of representations and affect the quality of human decisions.

Our study suggests that fruitful and robust human-algorithm interaction requires a reconsideration of what constitutes "communication" between the algorithm and the human. Epistemically and communicatively, it is not sufficient to extract causal representations effected by algorithms. Effective communication depends on understanding the whys of a causal representation effected by the algorithm — the standpoint from which it is generated — as well as reciprocity, negotiability and the ability to refer to a shared objective reality. We offer a set of desiderata for human-algorithm interaction in Discussion.

2 THEORY AND RELATED WORK

2.1 Human-algorithm Collaboration

One literature stream on human-algorithm interaction adopts a managerial or engineering perspective and takes performance as the object of study. What matter is not the epistemic content of human-algorithm interaction but the effect it has on performance, evaluated against some metric. In this vein, Fügener et al. analyze how human-algorithm interaction can lead to the "cyborgization" of human thought. This results in the loss of unique human knowledge that can contribute to effective decision-making. Fügener et al. show that humans and AI working together can outperform AI alone when the latter delegates to the former. Sun et al. study human deviations from algorithmic prescriptions in warehouse operations. They devise a machine learning algorithm to predict the deviations and show that incorporating them in logistics planning improves performance. Results from Kawaguchi's field experiment show that the failure to adopt algorithmic recommendation can lead to a gap between the nominal and the actual performance of the algorithm. Gao et al. develop a method that uses learning from bandit feedback to optimize human-AI collaboration.

The common thread behind this body of work is the conception of human-algorithm collaboration as a process that can and should be optimized in order to improve performance. To the extent that the process produces observable output in experiments, simulations, and real-world deployment, the data are used to understand decision quality and how it can be enhanced. This is an intellectual viewpoint that focuses on the "external" — measurements whose variances are related to observed variances in the structure of human-algorithm interaction. What is relatively less theorized is the epistemic content of this interaction. That is, beyond the fact that human knowledge and human actions can affect performance, how do humans process algorithmic predictions or recommendations *as information*?

2.2 Affective States

Another stream of literature focuses on internal psychological states induced by interaction with the algorithm. While the implicit goal might still be the improvement of a performance measure such as adoption, this body of work seeks to explain engagement psychologically or develop a normative framework for judging the conditions under which an internal psychological state is desirable. The literature has identified chiefly two affective states as important for human-algorithm interaction — trust and aversion. Dietvorst et al. introduce the concept of "algorithm aversion" and Dietvorst et al. show that agency helps to attenuate it. Lebovitz et al. investigate how in medical diagnosis the opacity of AI diagnostics can lead to a loss of trust via an increase in epistemic uncertainty. [Jacovi et al. 2021] formalize the notions of trust and trustworthiness in human-AI interaction and examine the criteria under which trust is normatively warranted. Glikson and Woolley review tangibility, transparency, reliability, and immediacy as factors that help to

inculcate cognitive trust and anthropomorphism as a factor that helps to inculcate emotional trust in the AI. Lastly, [Ullman and Malle 2018] develop a multidimensional measure of trust in the context of robotics.

What underlies this body of work is an "internal" view that psychological states are determined by how humans interface with the algorithm and in turn drive aspects of human-algorithm interaction. Furthermore, there is a normative dimension in so far as certain psychological states such as trust are only warranted under specific conditions.

2.3 The Duet of Representations

While an affective state clearly disposes the human towards a particular set of actions vis-à-vis the algorithm, it lacks the adaptive rationality that allows agents to respond to changes. Feelings of trust and aversion, once developed, are relatively constant, at least on the timescale of human-algorithm interaction. They do not explain individual instances of interaction. Consider interpersonal dynamics. One might be inclined to accept suggestions from a trustworthy friend, with such inclination assumed to be relatively stable over time. There will nonetheless be variability in one's actions if one is not to surrender one's agency completely. If agency is the ability to perform a difference-making action (e.g. accept or reject a suggestion) as it pertains to one's goal, then affective states alone cannot account for the differences in actions. What enables difference-making actions and agency is representation. Humans build mental models of the world and also attribute mental models — via theory of mind — to other agents [Lagnado 2021]. These models are often causal representations with interventionist[1] or counterfactual[2] semantics [Pearl and Mackenzie 2018; Woodward 2021]. Human collaboration requires, in addition to a representation of the shared goal or task, representations of other agents — more precisely, representations of other agents' representations of the task [Xiang et al. 2022]. When the human and the algorithm cooperate on a predictive task, it is the human's representation and the representation that she attributes to the algorithm that jointly enable the human to exercise her agency and perform difference-making actions.

Formally, the human constructs a representation R_h with respect to the predictive task from the space of human-interpretable representations $\mathcal{R}$. This representation could be a causal model with interventionist or counterfactual semantics [Pearl and Mackenzie 2018; Woodward 2021]. It could also be a simpler heuristic [Gigerenzer and Goldstein 1996; Gigerenzer et al. 1999]. For example, the representation for a binary classification task might be a sparse set of feature values that contributes to a negative prediction: $\{U = u_0\} \vee \{V = v_0\} \vee \{W = w_0\} \mapsto -1$. Since this representation defines the human's state of knowledge before using the algorithm, it can also be understood as the prior belief. The algorithm also effects a causal representation that relates features and labels. For example, a causal representation (that the human attributes to the algorithm) for a binary classification task might be that certain features values cause a positive prediction whereas others cause a negative prediction:: $\{U = u_0\} \vee \{V = v_0\} \vee \{W = w_0\} \mapsto 1; \{X = x_1\} \vee \{Y = y_1\} \vee \{Z = z_1\} \mapsto -1$.

Given an instance from the input space $\mathcal{I}$, the representation the algorithm effects and the human prior belief interact to induce a human action from the action space $\mathcal{A}$: $\mathcal{I} \times \mathcal{R} \times \mathcal{R} \mapsto \mathcal{A}$. Using the binary classification example from above, consider the input instance $\{U = u_0\}$ for which the algorithm gives a positive prediction. The causal representation attributed to the algorithm $R_a : \{U = u_0\} \mapsto 1$ conflicts with the prior belief $R_h : \{U = u_0\} \mapsto -1$. This might prompt the human to reject or revise the algorithm's prediction. The set of possible actions as well as the exact process by which an action is selected will vary by the predictive task. Bayesian updating [Bundorf et al. 2019], for example, might be appropriate for continuous labels.

[1] Setting X to x_0, Y would be y_0
[2] Had X been x_0, Y would have been y_0

2.4 Explanations as Compressed Representations

[Simon 2013] defines communication as "any process whereby decisional premises are transmitted from one member of an organization to another." Effective collaboration requires that the agents' states of knowledge be commensurate with the decision-making process. Sometimes routines and procedures encode past learning and constrain the agents to cooperate [Cyert et al. 1963]. At other times explicit communication between the agents is needed to establish an adequate basis for action. With respect to predictive tasks where the human shares in decisional authority and responsibility, the latter means that the algorithm's representation would have to be communicated to the human. Explanations can render complex representations human-interpretable and help to close the gap in shared knowledge of decisional premises.

In this capacity, explanations can be regarded as compressed representations: $R_a = \psi(h)$, where $|R_a| < |h|$. That is, an explanatory method $\psi(\cdot)$ extracts a compressed representation R_a of the original function $h$learned by the algorithm from the space of human-interpretable representations $\mathcal{R}$. The compressed representation R_a is more sparse than the underlying function h in some sense, e.g. the number of features. For example, LIME approximate the underlying model locally with sparse linear representations [Ribeiro et al. 2016]. SHAP model features as players in a cooperative game and extract the most relevant ones as explanations [Lundberg and Lee 2017]. Both have human-interpretable semantics — the former in the form of a sparse linear model, the latter in the form of a sparse set of relevant features.

Compressed representations generated by explanatory methods might also conflict with human priors. Explanations, however, increase the *salience* of any conflict with human priors by commanding cognitive attention and directing it to sparse and cogent representations with clear semantics. The fact that an explanation is explicitly and concisely shown to the human — for SHAP it would be a set of relevant features — can make the disagreement more conspicuous. This exacerbates the conflict and can affect the quality of human decision-making.

Fel et al. take the view that explanations help human users to learn to meta-predict model predictions. While achieving an understanding of the model sufficient for meta-predicting its predictions is a normative good that can be useful for many tasks, this conceptualization underplays the fact that in many real-world settings learning per se is not the express goal — action is. It cannot be assumed that a human user would suspend her prior beliefs in making decisions as she might when the objective is explicitly learning. An explanation, therefore, is not simply a piece of information to be used for improving one's understanding. It represents a distinct epistemic standpoint which can conflict with that of the human user to an extent that is consequential for actions.

2.5 Psychological Salience and Information Processing

An explanation directs attention to a sparse representation with clear semantics. This increases the *salience* of features that form part of the explanation, especially if the human user has a strong prior belief on how they should be related to the label and this prior belief conflicts with what is implied by the explanation. For example, the human might hold the belief that $\{U = u_0\} \mapsto -1$ and an explanation that $\{U = u_0\}$ contributes to a positive prediction would put the conflict in the crosshairs.

Research in psychology has shown that salience can have a large impact on human judgment [Taylor et al. 1979; Taylor and Fiske 1978] and even affect causal attribution [Fiske et al. 1982; Taylor and Fiske 1975]. In the context of machine learning explanations, the focus of attention on a sparse set of features for which the human user holds strong prior beliefs can induce overconfidence [Wan et al. 2022]. By directing attention to features on which the human user has a conflicting prior belief, the explanation also directs attention *away* from other features which could have been

part of the human user's information processing. This leads to causal overattribution and can affect the quality of human decision-making.

3 FIELD EXPERIMENT

The empirical context is a public employment service (PES) in the European Union. The PES provides services such as job referral and vocational training to unemployed individuals. When an individual becomes unemployed, she has to register at PES to receive financial support from the government. During the registration process, usually done in person, the registrant gives her data to a counselor who will review her case and support her in finding employment. According to an internal regulation, a counselor at the PES is obliged to assess the unemployed candidate's risk of LTU (long-term unemployment) upon registration, where LTU is defined as being involuntarily unemployed for a year or more.

We trained and implemented an XGBoost classification model that took as input candidate features and returned as output a raw probability score (risk score) and a risk assessment. Raw probability scores of LTU produced by XGBoost were converted into risk assessments of low, medium, and high, where high means a high probability of LTU. A risk assessment of high is equivalent to a positive prediction of LTU whereas a risk assessment of medium or low is equivalent to a negative prediction of LTU.

To explore the effect of explanations on human-algorithm interaction, we ran a field experiment from October 2019 to June 2020. The assignment of treatment was randomized at the level of job centers. Six centers were selected for the experiment, three for the treatment and three for the control group. Within a job center, candidates were assigned counselors available at the time of registration. After running the model, counselors were shown a risk assessment of low, medium or high and the raw probability score (risk score). The treatment group of counselors was additionally shown SHAP which comprised a set of six features. For a high (low) risk assessment, the top six features that increased (decreased) the probability of LTU were displayed; for a medium risk assessment, the top three features that, respectively, increased and decreased the probability of LTU were displayed. The counselors had the decisional authority and could either retain the algorithm's assessment or replace it with their own. They were also asked to rate their confidence in the final assessment on a Likert scale of 1 to 5. Data on the realized LTU outcomes of the candidates were collected in December 2021. Further information on the empirical setting can be found in Appendix A.

4 METHODS

4.1 Identifying the Conflict

To extract a sparse causal representation effected by the algorithm, we regress the algorithm's LTU prediction r on candidate features using LASSO logistic regression with ten-fold cross-validation for the control group. The regression yields sparse linear models $p(r = i) = \mathbf{x}^\intercal \boldsymbol{\beta}_i$, where $p(r = i)$ is the probability of the prediction being i (positive or negative), $\boldsymbol{\beta}_i$ the set of coefficients associated with prediction i, and $\mathbf{x}$ the sparse set of features significant for driving variances in the algorithm's LTU predictions.

Of the nine features, age, number of registrations, number of subsidy suspensions, and unemployment length are numeric variables. The rest are dummy variables derived from categorical variables. The representation attributed to the algorithm has the following possible semantics: if the unemployed candidate left her previous employment by mutual agreement, is college-educated, receives social integration subsidy, is in the age group above 56, is older, has had a fewer number of registrations, has had a fewer number of subsidy suspensions, and/or has been unemployed

Table 1. LASSO regression coefficients for the representation attributed to the algorithm

	LTU prediction	non-LTU	LTU
1	reason = contract ended	+0.304	-0.304
2	reason = mutual agreement	-0.347	+0.347
3	education = college	-0.127	+0.127
4	social integration subsidy = true	-1.163	+1.163
5	age	-0.133	+0.133
6	number of registrations	+0.037	-0.037
7	number of subsidy suspensions	+0.012	-0.012
8	age group : >56	-0.518	+0.518
9	unemployment length	-0.023	+0.023

for longer, then she is more likely to be (judged by the algorithm to be) in long-term unemployment. However, if the unemployed candidate left her previous employment because the contract ended, is younger, has had a greater number of registrations, has had a greater number of subsidy suspensions, and/or has been unemployed for shorter, then she is less likely to be (judged by the algorithm to be) in long-term unemployment.

R_a :

{reason = mutual agreement} $\vee$ {education = college}$\vee$

{social integration subsidy = true} $\vee$ {age group : >56} $\vee$ {age +}$\vee$

{number of registrations -} $\vee$ {number of subsidy suspensions -}$\vee$

{unemployment length +} $\mapsto$ LTU;

{reason = contract ended} $\vee$ {age -} $\vee$ {number of registrations +}$\vee$

{number of subsidy suspensions +} $\vee$ {unemployment length -} $\mapsto$ non-LTU.

The counselor either retains or adjusts the algorithm's risk assessment. We construct a variable a, with actions $a = -1$ (adjusting the algorithm's risk assessment downward), $a = 0$ (retaining the algorithm's risk assessment as it is), and $a = 1$ (adjusting the algorithm's risk assessment upward). To identify features where the counselor have a strong prior belief, we regress a on candidate features using LASSO multinomial regression with ten-fold cross-validation for the control group. This yields sparse linear models $p(a = j) = \mathbf{w}^\top \boldsymbol{\alpha}_j$, where $p(a = j)$ is the probability of action j, $\boldsymbol{\alpha}_j$ the set of coefficients associated with action j, and $\mathbf{w}$ the sparse set of features likely, ceteris paribus, to induce the counselors to take a particular action. The regression is run over the counselors as an aggregate so the prior belief is assumed to be held collectively. Both the sparsity and the linearity dovetail with the inductive bias for human mental models, which tend to be sparse and linear [Lombrozo 2007]

Except for age and number of interventions in job training, which are numeric variables, all are dummy variables derived from categorical variables. The human prior belief has the following semantics: if the unemployed candidate left her previous employment because she was a student, is college-educated, desires to find a job in scientific research, has a personal employment plan, and/or has had a greater number of interventions in job training, then she is less likely to be in long-term unemployment. However, if the candidate does not have EU/EEA nationality and/or has had a

Table 2. LASSO regression coefficients for the human prior belief

	action	down	same	up
1	reason = was student	+0.099	-0.061	-0.037
2	nationality = non-EU/EEA	-0.007	-0.027	+0.034
3	education = college	+0.464	-0.075	-0.389
4	age	-0.001	0.004	-0.004
5	desired job = scientific research	+0.012	-0.005	-0.008
6	personal employment plan = true	+0.196	-0.131	-0.065
7	prior personal employment plan = true	-0.022	+0.034	-0.012
8	number of interventions in job training	+0.061	-0.034	-0.027
9	was LTU = true	+0.052	+0.064	-0.116

fewer number of interventions in job training, then she is more likely to be in long-term unemployment. The features age, prior personal employment plan = true, and was LTU = true have positive coefficients for $a = 0$ (retaining the algorithm's prediction). Thus there isn't any strongly conflicting prior belief associated with them, although was LTU = true is slightly ambiguous as it also has a positive coefficient for downward adjustment.

R_h :

{reason = was student} $\vee$ {education = college} $\vee$ {desired job = scientific research}$\vee$

{personal employment plan = true} $\vee$ {number of interventions in job training + } $\mapsto$ non-LTU;

{nationality = non-EU/EEA} $\vee$ {number of interventions in job training -} $\mapsto$ LTU.

Comparing R_a and R_h, we identify college education as the feature where the human prior belief conflicts with the causal representation effected by the algorithm as a statistical regularity. Whereas the algorithm effects a sparse causal representation college education to a higher risk of long-term unemployment, the human prior belief seems to be of the opposite view — that college education should lead to a lower risk of long-term unemployment.

4.2 Identifying the Effects of Explanations on Decision Quality and Confidence

We first partition the experimental data into instances for which the algorithm gives a positive prediction of LTU (a risk assessment of high) and instances for which the algorithm gives a negative prediction of LTU (a risk assessment of medium or low). This is done for two reasons. Firstly, a conflict is specified only for a given algorithmic prediction. Secondly, the covariates and dependent variables in our regressions have different distributions for the two subsets of observations.

We construct a dummy variable Conflict to indicate the presence or absence of features where R_a and R_h conflict with each other in the explanation (shown to the treatment but not the control group). We estimate the following equation as our first model using logistic regression:

$$\begin{aligned}\text{Decision Quality} = \beta_0 + \beta_1 \text{Exposed} + \beta_2 \text{Conflict} + \beta_3 \text{Exposed} \times \text{Conflict} \\ + \beta_4 \text{Risk Score} + \text{time fixed effects} + \varepsilon. \end{aligned} \quad (1)$$

where Decision Quality is the accuracy of the final assessment and Exposted indicates treatment status. We use Risk Score, which is native to XGBoost, as a control variable that stratifies the data instances into bins of predictions of equal difficulty. The sign and statistical significance of β_3 tell us whether showing features on which the counselors hold a conflicting prior belief as part of the explanation affects the decision quality.

A second model, similarly estimated using logistic regression, explores the heterogeneity of the mechanism by adding Adjustment as an interaction variable, where Adjustment is a dummy variable indicating whether the counselor has adjusted the algorithm's LTU prediction:

$$\begin{aligned}\text{Decision Quality} = \beta_0 &+ \beta_1 \text{Exposed} + \beta_2 \text{Conflict} + \beta_3 \text{Adjustment} \\ &+ \beta_4 \text{Exposed} \times \text{Conflict} + \beta_5 \text{Exposed} \times \text{Adjustment} + \beta_6 \text{Conflict} \times \text{Adjustment} \\ &+ \beta_7 \text{Exposed} \times \text{Conflict} \times \text{Adjustment} + \beta_8 \text{Risk Score} + \text{time fixed effects} + \varepsilon. \end{aligned} \tag{2}$$

The sign and statistical significance of β_4 (β_7) tell us whether showing features on which the counselors hold a conflicting prior belief as part of the explanation affects the decision quality, when the counselor retains (adjusts) the algorithm's prediction.

In a third model we examine the impact of explanations on counselors' confidence in the final assessment using linear regression:

$$\begin{aligned}\text{Confidence} = \beta_0 &+ \beta_1 \text{Exposed} + \beta_2 \text{Conflict} + \beta_3 \text{Adjustment} \\ &+ \beta_4 \text{Exposed} \times \text{Conflict} + \beta_5 \text{Exposed} \times \text{Adjustment} + \beta_6 \text{Conflict} \times \text{Adjustment} \\ &+ \beta_7 \text{Exposed} \times \text{Conflict} \times \text{Adjustment} + \beta_8 \text{Risk Score} + \text{time fixed effects} + \varepsilon. \end{aligned} \tag{3}$$

The sign and statistical significance of β_4 (β_7) tell us whether showing features on which the counselors hold a conflicting prior belief as part of the explanation increases or decreases confidence, when the counselor retains (adjusts) the algorithm's prediction.

5 RESULTS

For model 1, Conflict is equal to 1 when college education is displayed as part of the explanation (or would have been displayed to the control group as part of the explanation had their treatment condition been different). Only the subset of instances for which the algorithm gives a positive prediction of LTU is included. The regression results show that a positive prediction with a higher Risk Score is more likely to be correct. The coefficient for Exposed × Conflict is negative and statistically significant. This means that displaying college education as part of the explanation degrades the quality of decision-making.

Regression results for model 2 show that there is heterogeneity in the effects of explanations on decision quality with the larger part of the decrease coming from when counselors adjust the algorithm's positive prediction of LTU, as indicated by the negative and statistically significant coefficient for Exposed × Conflict × Adjustment.

Finally, regression results for model 3 show that displaying college education as part of the explanation has polarizing effects on confidence. It reduces confusion and increases confidence when counselors retain the algorithm's prediction, as indicated by the positive and statistically significant coefficient for Exposed × Conflict. On the other hand it draws attention to the conflict and decreases confidence when counselors adjust the algorithm's prediction, as indicated by the negative and statistically significant coefficient for Exposed × Conflict × Adjustment.

	Dependent variable:		
	Decision Quality	Decision Quality	Confidence
	(1)	(2)	(3)
Exposed	−0.152*	−0.125	-0.006
	(0.064)	(0.083)	(0.018)
Exposed × Conflict	−0.439***	−0.231	0.171***
	(0.129)	(0.187)	(0.041)
Exposed × Conflict × Adjustment		−0.662*	−0.331***
		(0.275)	(0.063)
Risk Score	3.826***	2.395***	0.164
	(0.397)	(0.419)	(0.089)
2020Q1	0.362***	0.396***	−0.030
	(0.068)	(0.072)	(0.016)
2020Q2	0.326***	0.410***	−0.199***
	(0.071)	(0.076)	(0.017)
Observations	5,728	5,728	5,728
Log-likelihood	-3743.969	-3411.2	
R^2			0.048
Adjusted R^2			0.046
Note:			*p<0.05; **p<0.01; ***p<0.001

6 DISCUSSION

Conflict between epistemic standpoints defines the kind of rationality that enables actions. If there is no *epistemic conflict*, agency — the ability to perform a difference-making action as it pertains to one's goal — would not be realized.

In our field experiment, the conflict between the human's prior belief that college education is negatively associated with long-term unemployment and the obverse representation attributed to the algorithm leads to actions that worsen the quality of decision-making for the time period during which the pilot was run. An alternative scenario, however, can be conceived where the human's prior beliefs encode useful information about the world that the algorithm is not privy to. In such a scenario, epistemic conflict can improve the quality of decision-making. Nevertheless it is not possible for an organization to determine ex-ante whether epistemic conflict would degrade or enhance the quality of decision-making.

Explanations perform an epistemic or communicative function by rendering complex representations human-interpretable. This however does not resolve the conflict of representations and may in fact — because explanations direct attention to cogent representations with explicit semantics — exacerbate it. As can be seen with positive predictions of LTU in our field experiment, showing college education as part of the explanation degrades the quality of decision-making.

Bare communication of representations, therefore, is not sufficient. Such communication is oddly solipsistic in that each representation is committed to its own epistemic standpoint. What should drive human-algorithm interaction is not the mere fact of epistemic conflict but its *whys*, just as human beings do not just insist on their differences but act to understand and bridge them. A more extensive communicative rationality is needed to enable actions that would improve the quality of decision-making.

We believe there are four desiderata for such communicative rationality. The first desideratum is *understanding*. The human should understand the algorithm's epistemic standpoint. In our empirical context, this means being imparted information about model training as well as possible reasons for certain representations (e.g. *why* college education is associated with higher risk of long-term unemployment). The second desideratum is *reciprocity*. Human-AI interaction tends to be unidirectional. Increasing the algorithm's understanding of the human's epistemic standpoint and prior beliefs can improve communication. The third desideratum is *negotiability*. Reciprocal understanding facilitates negotiation where arguments can be developed, evidences arrayed, biases identified, and confidences gauged. The last desideratum is a *shared reality*. In our empirical context, the counselors do not receive feedback on the accuracy of their judgment. Convergence to a shared reality is possible if the counselors are made aware of where each of the two parties has erred.

REFERENCES

[1] Kate Bundorf, Maria Polyakova, and Ming Tai-Seale. 2019. *How do humans interact with algorithms? Experimental evidence from health insurance.* Technical Report. National Bureau of Economic Research.

[2] Richard M Cyert, James G March, et al. 1963. *A behavioral theory of the firm.* Vol. 2. Englewood Cliffs, NJ.

[3] Manuel DeLanda. 2021. *Materialist Phenomenology: A Philosophy of Perception.* Bloomsbury Publishing.

[4] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[5] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[6] Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. 2021. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. *arXiv preprint arXiv:2112.04417* (2021).

[7] Susan T Fiske, David A Kenny, and Shelley E Taylor. 1982. Structural models for the mediation of salience effects on attribution. *Journal of Experimental Social Psychology* 18, 2 (1982), 105–127.

[8] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2021. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly (MISQ)-Vol* 45 (2021).

[9] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2022. Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. *Information Systems Research* 33, 2 (2022), 678–696.

[10] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-ai collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614* (2021).

[11] Gerd Gigerenzer and Daniel G Goldstein. 1996. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review* 103, 4 (1996), 650.

[12] Gerd Gigerenzer, PM Todd, and ABC Group. 1999. Simple Heuristics that Make us Smart. (1999).

[13] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[14] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 624–635.

[15] Kohei Kawaguchi. 2021. When will workers follow an algorithm? A field experiment with a retail business. *Management Science* 67, 3 (2021), 1670–1695.

[16] David A. Lagnado. 2021. *Explaining the Evidence: How the Mind Investigates the World.* Cambridge University Press. https://doi.org/10.1017/9780511794520

[17] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 2022. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science* (2022).

[18] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology* 55, 3 (2007), 232–257.

[19] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems.* 4765–4774.

[20] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect.* Basic books.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 1135–1144.

[22] Herbert A Simon. 2013. *Administrative behavior.* Simon and Schuster.

[23] Jiankun Sun, 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 (2022), 846–865.

[24] Shelley E Taylor, Jennifer Crocker, Susan T Fiske, Merle Sprinzen, and Joachim D Winkler. 1979. The generalizability of salience effects. *Journal of Personality and Social Psychology* 37, 3 (1979), 357.

[25] Shelley E Taylor and Susan T Fiske. 1975. Point of view and perceptions of causality. *Journal of Personality and Social Psychology* 32, 3 (1975), 439.

[26] Shelley E Taylor and Susan T Fiske. 1978. Salience, attention, and attribution: Top of the head phenomena. In *Advances in experimental social psychology*. Vol. 11. Elsevier, 249–288.

[27] Daniel Ullman and Bertram F Malle. 2018. What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In *Companion of the 2018 acm/ieee international conference on human-robot interaction*. 263–264.

[28] Charles Wan, Rodrigo Belo, and Leid Zejnilovic. 2022. Explainability's Gain is Optimality's Loss? How Explanations Bias Decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 778–787.

[29] James Woodward. 2021. *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.

[30] Yang Xiang, Natalia Vélez, and Samuel J Gershman. 2022. Collaborative decision making is grounded in representations of other people's competence and effort. (2022).

A EMPIRICAL SETTING

Table 3. Treatment assignment to job centers

	pre-pilot		pilot		
job center	registrations	appointments/month	registrations	appointments/month	treatment status
1	11958	213	13139	169	0
2	9406	191	10263	160	1
3	3396	99	3743	72	0
4	5717	110	6022	88	1
5	3889	78	4379	69	0
6	7016	135	7336	100	1

B FULL REGRESSION RESULTS

	Dependent variable:		
	Decision Quality	Decision Quality	Confidence
	(1)	(2)	(3)
Exposed	−0.152*	−0.125	-0.006
	(0.064)	(0.083)	(0.018)
Exposed × Conflict	−0.439***	−0.231	0.171***
	(0.129)	(0.187)	(0.041)
Exposed × Conflict × Adjustment		−0.662*	−0.331***
		(0.275)	(0.063)
Conflict	0.135	−0.050	−0.100**
	(0.100)	(0.144)	(0.031)
Adjustment		−1.887***	−0.191***
		(0.119)	(0.027)
Exposed × Adjustment		0.230	0.108**
		(0.153)	(0.034)
Conflict × Adjustment		1.179***	0.349***
		(0.213)	(0.048)
Risk Score	3.826***	2.395***	0.164
	(0.397)	(0.419)	(0.089)
2020Q1	0.362***	0.396***	−0.030
	(0.068)	(0.072)	(0.016)
2020Q2	0.326***	0.410***	−0.199***
	(0.071)	(0.076)	(0.017)
Observations	5,728	5,728	5,728
Log-likelihood	-3743.969	-3411.2	
R^2			0.048
Adjusted R^2			0.046
Note:			*p<0.05; **p<0.01; ***p<0.001