In the Eye of the Beholder: Robust Prediction with Causal User Modeling

Anonymous Authors[1]

Abstract

Accurately predicting the relevance of items to users is crucial to the success of many social platforms. Conventional approaches train models on logged historical data; but recommendation systems, media services, and online marketplaces all exhibit a constant influx of new content—making relevancy a moving target, to which standard predictive models are not robust. In this paper, we propose a learning framework for relevance prediction that is robust to changes in the data distribution. Our key observation is that robustness can be obtained by accounting for *how users causally perceive the environment*. We model users as boundedly-rational decision makers whose causal beliefs are encoded by a causal graph, and show how minimal information regarding the graph can be used to contend with distributional changes. Experiments in multiple settings demonstrate the effectiveness of our approach.

1. Introduction

Across a multitude of domains and applications, machine learning has become imperative for guiding human users in many of the decisions they make [41,53,10]. From recommendation systems and search engines to e-commerce platforms and online marketplaces, learned models are regularly used to filter content, rank items, and display select information—all with the primary intent of helping users choose items that are relevant to them. The predominant approach for learning in these tasks is to train models to accurately predict the relevance of items to users, but since training is often carried out on logged historical records, even highly-accurate models remain calibrated to the distribution of *previously* observed data on which they were trained [7,54,51]. Given that in virtually any online platform the distribution of content naturally varies over time and location—due to trends and fashions, innovation, or forces of supply and demand—models trained on logged data may fail to correctly predict the choices and preferences of users on unseen, future distributions [12,1,18,26,33].

In this paper, we present a novel conceptual framework for learning predictive models of user-item relevance that are robust to changes in the underlying data distribution. Our approach is built around two key observations: (i) that relevance to users is determined by the way in which users *perceive* value, and (ii) that this process of value attribution is *causal* in nature. As an example, consider a video streaming service in which a user u is trying to determine whether watching a certain movie will be worthwhile. To make this decision, $y \in \{0, 1\}$, the user has at her disposal a feature description of the movie, x, and a system-generated, personalized relevance score, r (e.g., "a 92% match!"). How will she integrate these two informational sources into a decision? We argue that this crucially hinges on her belief as to *why* a particular relevance score is coupled to a particular movie. For example, if a movie boasts a high relevance score, then she might suppose this score was given *because* the system believes the user would like this movie. Another user, however, may reason differently, and instead believe that high relevance scores are given *because* movies are sponsored; if she suspects this to be a likely scenario, her reasoning should have a stark effect on her choices. In both cases, perceived values stem from how each user causally interprets the recommendation environment, and the underlying causal structure determines how belief regarding value changes, in response to changes in important variables.

Here, we show how knowledge regarding the causal perceptions of users can be leveraged for providing distributional robustness in learning. A primary concern for robust learning is the reliance of predictions on spurious correlations [3]; here we argue that spuriousness can *result* from causal perceptions underlying user choice behavior. To see the relation between causal perceptions and spuriousness, assume that in our movies example above, the training data exhibits a strong correlation between users' choices of movies, y, and a 'genre' feature, x_g. A predictive model optimized for accuracy will likely exploit this association, and rely on x_g for prediction. Now, further assume that what *really*

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

drives user satisfaction is 'production quality', x_q; if x_g and x_q are *spuriously* correlated in the training data, then once the distribution of genres naturally changes over time, the predictive model can fail: the association between x_g and y, on which predictions rely, may no longer hold.

In essence, our approach casts robust prediction of personalized relevance as a problem of out-of-distribution (OOD) learning, but carefully tailored to settings where data generation is governed by users' causal beliefs and corresponding behavior. There is a growing recognition of how a causal understanding of the learning environment can improve cross-domain generalization [3,50]; our key conceptual contribution is the observation that, in relevance prediction, users' perceptions *are* the causal environment. Thus, there is no 'true' causal graph—all is in the eye of the beholder. To cope with this, we model users as reasoning about decisions through a causal graph [34,43]— allowing our approach to anticipate how changes in the data translate to user behavior.

Building on this idea, as well as on recent advances in the use of causal modeling for out-of-distribution learning [3, 50, 49], we show how various levels of knowledge regarding users' causal beliefs can be utilized for learning distributionally robust predictive models. To encourage predictions $\hat{y}$ to be invariant to changes in the environment e, our approach enforces independence between $\hat{y}$ and e. This is achieved with regularization [21]. In general, different graphs require different regularization [49]; unfortunately, inferring user graphs is likely to be costly and challenging. Our key observation is that, for user graphs, it suffices to know which of two *classes* the graph belongs to—*causal* or *anti-causal*— determined by the direction of edges between x, r and y (Fig. 1a). The user's class, which is easier to infer, determines the regularization scheme.

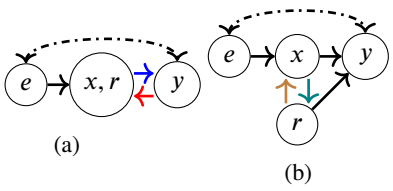

Figure 1: Simplified graphs describing users of different: (a) *classes*: *causal* or *anti-causal*, and (b) *sub-classes*: *believer* or *skeptic* (here shown for a causal user). Dashed lines indicate possible spuriousness (e.g., via selection).

Nonetheless, more fine-grained information can still be useful. We show the following novel result: if two users generate the same data, but differ in their underlying graph, they will have different optimal *out-of-distribution* models. The reason for this is that, to achieve robustness, regularizing for independence will result in the discarding of different information for each user. Operationally, this means that learning should include different models for each user-type. Here again we show that minimal additional information is useful, and focus on subclasses of graphs that differ only in the direction of the edges between x and r (Fig. 1b); nonetheless, our result applies broadly and may be of general interest for causal learning. In Appendix A we present a thorough empirical evaluation of our approach, where we explore the benefits of different forms of knowledge regarding users' causal beliefs: whether they are *casual* or *anti-causal*, and the directionality of $x \leftrightarrows r$. Importantly, our results also imply that *not* accounting for causal aspects of user decision-making, can result in poor OOD performance.

2. Modelling Approach

2.1. Learning Setting

In our setting, data consists of users, items, and choices. Users are described by features $u \in \mathbb{R}^{d_u}$, and items are described by two types of features: intrinsic item properties, $x \in \mathbb{R}^{d_x}$ (e.g., movie genre, plot synopsis, cast and crew), and information provided by the platform, $r \in \mathbb{R}^{d_r}$ (e.g., recommendation score, user reviews). We will sometimes make a distinction between features that are available to users, and those that are not; in such cases, we denote unobserved features by $\bar{x}$, and with slight abuse of notation, use x for the remaining observed features (we assume r is always observed). Choices $y \in \{0, 1\}$ indicate whether a user u chose to interact with a certain item (x, r).

As we are interested in robustness to distributional change, we follow the general setup of *domain generalization* [6,25,5,50] in which there is a collection of environments, denoted by a set $\mathcal{E}$, and each environment $e \in \mathcal{E}$ defines a different joint distribution D^e over (u, x, r, y). We assume there is training data available from a subset of K environments, $\mathcal{E}_{\text{train}} = \{e_1, \ldots, e_K\} \subset \mathcal{E}$, with datasets $S_k = \{(u_{ki}, x_{ki}, r_{ki}, y_{ki})\}_{i=1}^{m_k}$ drawn i.i.d from the corresponding D^{e_k}. We denote the pooled training distribution by $D_{\text{train}} = \cup_{e \in \mathcal{E}_{\text{train}}} D^e$ and the pooled training data by $S = \cup_k S_k$ with $m = \sum_k m_k$. Our goal is to learn a robust predictive model $\hat{y} = f(u, x, r; \theta) := f_u(x, r; \theta)$.

Robustness via causal graphs. The type of robustness that we would like our model to satisfy is *counterfactual invariance* (CI) [49]. Denoting $x(e), r(e)$ as the counterfactual features that would have been observed had the environment been set to e, we define a model f_u as CI if $\forall e, e' \in \mathcal{E}$ it holds a.e. that $f_u(x(e'), r(e'); \theta) = f_u(x(e), r(e); \theta)$.

The challenge in obtaining CI predictors is that at train time we only observe a subset of the environments, $\mathcal{E}_{\text{train}} \subset \mathcal{E}$, while CI requires independence to hold for *all* environments $e \in \mathcal{E}$. To reason formally about the role of e in the data generating process, and hence about the type of distribution shifts under which our model should remain invariant, it is

common to assume that a causal structure underlies data generation [22,3]. This is often modeled as a (directed) causal graph [34]; robustness is then defined as insensitivity of the predictive model to changes (or 'interventions') in the variable e, which can trigger changes in other variables that lie 'downstream' in the graph. To encourage robustness, a common approach is to construct a learning objective that avoids spurious correlations by enforcing certain conditional independence relations to hold, e.g., via regularization (see §3). The question of *which* relations are required can be answered by examining the graph and the conditional independencies it encodes (between e, x, r, y, and $\hat{y}$). Unfortunately, inferring the causal graph is in general hard; however, determining the 'correct' learning objective may require only partial information regarding the graph.

2.2. Users as decision makers

Rational users. To see how modeling users as causal decision-makers can be helpful, consider first a conventional 'correlative' approach for training f_u, e.g. by minimizing the loss of a corresponding score function $v_u(x, r) = v(u, x, r)$, and predicting via $\hat{y} = \text{argmax}_{y\in\{0,1\}}\, y v_u(x, r) = \mathbb{1}\{v_u(x, r) > 0\}$. From the perspective of user modeling, v_u can be interpreted as a personalized 'value function'; this complies with classic Expected Utility Theory (EUT) [48], in which users are modeled as rational agents acting to maximize (expected) value, and under *full information*. From a causal perspective, this approach is equivalent to assuming a graph in which all paths from e to y are blocked by x, r—which is akin to assuming no spurious pathways, and so handicaps the ability to avoid them.

Boundedly-rational users. We propose to model users as boundedly-rational decision-makers, under the key assertion that users' decisions take place under inherent *uncertainty*. Uncertainty plays a key role in how we, as humans, decide: our actions follow not only from what we know, but also from how we account for what we don't know. Nonetheless, despite being central to most modern theories of decision making [24], explicit modeling of user-side uncertainty is currently rare within ML [36,2,37]. Our modeling approach acknowledges that users *know* some features are unobserved, and that this influences their actions. Consider a user shopping online for a vintage coat, and considering whether to buy a certain coat. The coat's description includes several intrinsic properties x as well as certain platform-selected information r The user wants to make an informed decision, but knows some important information, $\bar{x}$, is missing If she is concerned about buying a modern knockoff, how should she act? A common approach is to extend EUT to support uncertainty by modelling users as integrating subjective beliefs about unobserved variables, $p_u(\bar{x}|x, r, e)$, into a conditional estimate of value, $\tilde{v}_u(x, r|e)$, over which choices y are made: $\tilde{v}_u(x, r|e) = \sum_{\bar{x}} v_u(x, \bar{x}, r) p_u(\bar{x}|x, r, e)$

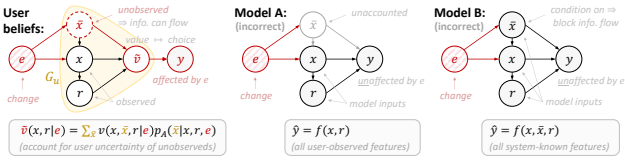

Figure 2: **(Left)** User causal beliefs. Integration of uncertainty in unobserved $\bar{x}$ results in a direct link $e \to y$ (note y is deterministic of $\tilde{v}$). This reveals a source of possible spuriousness, but also suggests how to treat it. **(Center)** A predictive model $f(x, r)$ using only observed features. Learning results in x, r compensating for e, and so f cannot account for change. **(Right)** A predictive model $f(x, \bar{x}, r)$ using all available features. Learning results in wrongly using variation in $\bar{x}$ to explain y.

where $y = \mathbb{1}\{\tilde{v}_u(x, r|e) > 0\}$. Here, $p_u(\bar{x}|x, r, e)$ describes a user's (probabilistic) belief regarding the conditional likelihood of each $\bar{x}$ and $v_u(x, \bar{x}, r)$ describes the item's value to the user *given* $\bar{x}$ Importantly, note that uncertainty beliefs $p_u(\bar{x}|x, r, e)$ can be environment-specific In turn, value estimates $\tilde{v}$ and choices y can also rely on e. For clarity, since y is a deterministic function of $\tilde{v}$, we will simply refer to y as a function of x, r, and e.

Causal user graphs. One interpretation of users' utility function is that they cope with uncertainty by employing *causal reasoning* [43], this aligning with a predominant approach in the cognitive sciences that views humans as acting based on 'mental causal models' [42]. Here we follow [43] and think of users as reasoning through personalized *user causal graphs*, denoted G_u. The structure of G_u expresses u's causal beliefs—namely which variables causally affect others—and its factors correspond to the conditional terms (p_u and v_u) in the utility function. A key modeling point is that users can vary in their causal perceptions; hence, different users may have different graphs that encode different conditional independencies, these inducing different simplifications of the conditional terms. For example, a user that believes movies with a five-star rating (r) are worthwhile regardless of their content (x) would have $v_u(x, \bar{x}, r)$ reduced to $v_u(\bar{x}, r)$, since $v \perp\!\!\!\perp x|r$; meanwhile, a user who, after reading a movie's description (x), is unaffected by its rating (r), would have $v_u(x, \bar{x})$ instead, since $v \perp\!\!\!\perp r|x$.

2.3. Towards robust learning

Recall that our goal is to learn a predictor f that is unaffected by spurious correlations, and that these can materialize through direct edges $e \to y$. Continuing our illustrative example, assume that the causal beliefs of our boundedly-rational user are encoded by the graph in Figure 2. The graph does not have a direct edge $e \to y$; however, by accounting for uncertainty (i.e., choosing via $\tilde{v}$), the user behaves 'as if' there actually was a direct edge—since integrating over $\bar{x}$ 'removes' it from the indirect path $e \to \bar{x} \to y$. This support

our main argument: by making decisions, users can *generate* spurious correlations in the data (Fig. 2 ('User model')).

The above has two implications. First, it shows how conventional approaches can fail. For example, since users observe only x and r, one reasonable approach is to discard $\bar{x}$ and train a predictor $f_u(x, r; \theta)$ in hopes of mimicking user choice behavior (Fig. 2 ('Model A')). In graphical terms, this means learning 'as if' there is no edge $e \to y$. But the reliance of users on $\bar{x}$ for producing y means that, effectively, such an edge exists. The result of this is that f_u can learn to use x and r to compensate for the constant effect of e on y. Conversely, the system may choose to learn using all available information, namely train $f_u(x, \bar{x}, r; \theta)$ (Fig. 2 ('Model B')). This make sense if the goal is in-distribution generalization; for OOD, this not only creates an illusion of separation between e and y, but also erroneously allows to use the variation in $\bar{x}$ to explain y. As a result, f_u will likely overfit to distributions in $\mathcal{E}_{\text{train}}$. Second, the awareness to how users account for uncertainty suggests a means to combat spuriousness. The utility function shows that y depends on x, r, and e; hence, since our goal is to discourage the dependence of $\hat{y}$ on e, it follows that (i) functionally, f_u should depend only on x and r but (ii) f_u should be learned in a way that controls for variation in e.

3. Learning With Causal User Models

Our approach to robust learning is based on regularized risk minimization, where regularization acts to discourage variation in predictions across environments [49,50]. Our learning objective is: $\text{argmin}_{f \in F}\, L(f; S) + \lambda R(f; S_1, \ldots, S_K)$ where R is a regularization term with coefficient λ. The role of R is to penalize f for violating certain statistical independencies; the question of *which* independencies should be targeted—and hence the precise form that R should have—can be answered by the underlying causal graph [49]. Knowing the full user graph can certainly help, but relying on this is impractical. Luckily, coarse information regarding the graph can be translated into necessary conditions for distributional robustness, and in App. A we show that these can go a long way towards learning robust models. We focus on two methods for promoting statistical independence of $\hat{y}$ and e [49]: MMD [21] and CORAL [47] (see App. E).

3.1. User graph classes: causal vs. anti-causal

Following our example in Fig. 1a, consider users of two types: a 'causal' user $u_{\to y}$ that believes value is an *effect* of an item's description (i.e., $D^e(x, r, y \mid u = u_{\to y})$ is entailed by the graph $x, r \to y$ for each $e \in \mathcal{E}$), and an 'anti-causal' user $u_{\leftarrow y}$ that believes the item's value *causes* its description (i.e., $D^e(x, r, y \mid u = u_{\leftarrow y})$ is entailed by $x, r \leftarrow y$ respectively). Our next result shows that: (i) $u_{\to y}$ and $u_{\leftarrow y}$ require *different* regularization schemes; but (ii) that the appropriate scheme is fully determined by their type—*irrespective* of any other properties of their graphs. Thus, from a learning perspective, it suffices to know which of two classes a user belongs to: causal, or anti-causal.

Proposition 1. *Let f be a CI model and assume y and e are confounded (e.g., $e \to y$ exists), then:*

(1) $f_{u_{\to y}}$ *must satisfy* $P_{D^e}\left(f_{u_{\to y}}(x, r)\right) = P_{D^{e'}}\left(f_{u_{\to y}}(x, r)\right)$

(2) $f_{u_{\leftarrow y}}$ *must satisfy*

$$P_{D^e}\left(f_{u_{\leftarrow y}}(x, r) \mid y\right) = P_{D^{e'}}\left(f_{u_{\leftarrow y}}(x, r) \mid y\right), y \in \{0, 1\}$$

On the other hand, $f_{u_{\to y}}$ need not necessarily satisfy (2), and $f_{u_{\leftarrow y}}$ need not necessarily satisfy (1).

If we fail to enforce these constraints during learning, then we will not learn a CI classifier. On the other hand, enforcing unnecessary constraints restricts our hypothesis class and hence limits performance. The proof follows directly from [49] (see App. B). The distinction between causal and anti-causal prescribes the appropriate regularization. For any user u, to encourage $f_u(x, r)$ to be invariant to changes in e, set:

$$R(f; S) = \begin{cases} \sum_k \text{MMD}(\Phi_{k,u}, \Phi_{-k,u}) & u \text{ is causal} \\ \sum_y \sum_k \text{MMD}(\Phi^{(y)}_{k,u}, \Phi^{(y)}_{-k,u}) & u \text{ is anti-causal} \end{cases}$$

where $\Phi_{k,u}, \Phi^{(y)}_k$ includes the subset of examples with user u and label y, respectively.

3.2. User graph subclasses: inter-feature relations

Consider now two users that are of the same class (i.e., causal or anti-causal), but perceive differently the causal relations between x and r: a *believer*, $u_{x \to r}$, who believes recommendations follow from the item's attributes; and a *skeptic*, $u_{x \leftarrow r}$, who presumes that the system reveals item attributes to match a desired recommendation (see Fig.1b). Our main result shows that even if both users share the same objective preferences—to be *optimally* invariant, each user may require her own, independently-trained model.

Proposition 2. *Let $u_{x \to r}, u_{x \leftarrow r}$ be two users of the same class (i.e., causal or anti-causal) but of a different subclass (i.e., believer and skeptic, respectively). Even if there is a single predictor f which is optimal for the pooled distribution D_{train}, each user can have a different optimal CI predictor.*

Proof is in Appendix B. Prop. 2 can be interpreted as follows: Take some u, and 'counterfactually' invert the edges between x and r. In some cases, this will have no effect on u's behavior under $\mathcal{E}_{\text{train}}$, and so any f that is optimal in one case will be optimal in the other. Nonetheless, for optimality to carry over to *other* environments—different predictors may be needed. This is since each causal structure implies a different interventional distribution, and hence a different set of CI predictors: e.g., in $G_{x \leftarrow r}$, the v-structure $e \to x \leftarrow r$ suggests that an invariant predictor may depend on r, yet in $G_{x \to r}$ it cannot. In §A.1, we empirically evaluate this.

4. Experiments and Results

4.1. Learning with *causal* users: text-based beer recommendation

Data. We use *RateBeer*, a dataset of beer reviews with over 3M entries and spanning $\sim$ 10 years [31]. We use the data to generate beer features x (e.g., popularity, average rating) and r (e.g., textual review embeddings) and user features u (e.g., average rating, word counts). Given a sample (u, x, r), our goal is to predict a (binarized) rating y. Here we focus on *causal* users, and so would like labels y to expresses causal user beliefs. The challenge is that our observational data is not necessarily such. To simulate causal user behavior, we rely on the observation that $x, r \rightarrow y$ means "changes in x, r affect y", and for each u create an individualized empirical distribution of 'counterfactual' samples (x', r', y') that approximate the entire intervention space (i.e., all counterfactual outcomes y' under possible interventions $(x, r) \mapsto (x', r')$). Training data is then generated by sampling from this space.

We consider each year as an environment e, with each e inducing a distribution over (u, x, r). We implement spuriousness via selection: Each e entails different fashionable 'tastes' in beer, expressed as a different weighting over the possible beer types (e.g., lager, ale, porter). Labels are then made to correlate with tastes in a certain temporal pattern. This serves as a mechanism for spurious correlation.

Results. Fig. 3 compares the performance over time of three training procedures that differ only in the type of regularization applied: *causal*, *anti-causal*, and *non-causal*. Our data includes behavior generated by causal-class users; results demonstrate the clear benefit of using a behaviorally-consistent regularization scheme (here, causal). Note the causal approach is not optimal in 2006 and 2008; this is since correlations in $e \leftrightarrow y$ are set to make these years similar to the training data. However, when tastes shift other approaches collapse, while the causal approach remains stable.

4.2. Learning with *anti-causal* users: clothing-style recommendation

Data. We use the *fashion product images* dataset, which includes includes 44.4k fashion items described by images, attributes, and text. Here we focus on *anti-causal* users, and generate data in a way similar to §4.1, but using an anti-causal intervention space. In this experiment we let user choices $y \in \{0, 1\}$ depend on an item's image and color, which can be either red or green; in this way, x is the item's grayscale image, and r its hue (which we control). Here we consider environments e that induce varying degrees of spurious correlations between color and user choices, $P(y = 1|\texttt{red}) = P(y = 0|\texttt{green}) = p_e$. For the test set we use $p_e = 0.8$, and experiment with training data that gradually deviate from this relation, i.e., having $p_{e'} \in [0.1, 0.8]$.

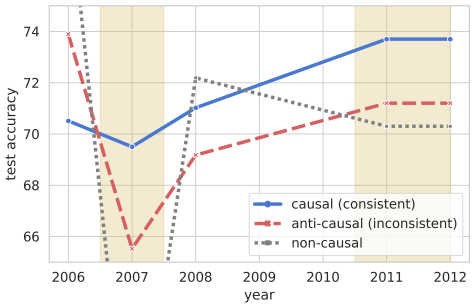

Figure 3: ***RecBeer* Results**. For each year, models are trained on past data (starting 2002), and predict on the following year. The *causal* training scheme, consistent with the user class, outperforms other methods when beer-type fashions (e) changes. Periods with substantial change are highlighted in tan.

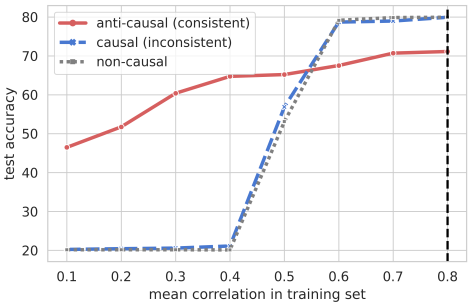

Figure 4: ***RecFashion* Results**. Environments vary in the correlation between item colors and user choices. The *anti-causal* regularization scheme, consistent with the user class, outperforms methods when test-time deviates from train-time correlation (=0.8). When correlations flip (< 0.5), other methods crash.

Results. Fig. 4 shows that consistent regularization (here, anti-causal) outperforms other alternatives whenever correlations deviate from those observed in training. Once correlations flip, both *causal* and *non-causal* approaches fail catastrophically; the *anti-causal* approach remains robust.

5. Discussion

Humans beings perceive the world causally; our paper argues that to cope with a world that *changes*, learning must take into account how humans believe these changes take effect. We identify one key reason: in making decisions under uncertainty, users can *cause* spurious correlations to appear in the data. Towards this, we propose to employ tools from invariant causal learning, but in a way that is tailored to how humans make decisions, this drawing on economic models of bounded-rationality. Our approach relies on regularization for achieving invariance, with our main point being that *how* and *what* to regularize can be derived from users' causal graphs.

References

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 42–46, 2017.

[2] Reut Apel, Ido Erev, Roi Reichart, and Moshe Tennenholtz. Predicting decisions in language based persuasion games. *Journal of Artificial Intelligence Research*, 73:1025–1091, 2022.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Alexis Bellot and Mihaela van der Schaar. Generalization and invariances in the presence of unobserved confounding. *arXiv preprint arXiv:2007.10653*, 2020.

[5] Eyal Ben-David, Nadav Oved, and Roi Reichart. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433, 2022.

[6] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24:2178–2186, 2011.

[7] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pages 104–112, 2018.

[8] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

[9] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[10] Alison Callahan and Nigam H Shah. Machine learning in healthcare. In *Key Advances in Clinical Informatics*, pages 279–291. Elsevier, 2017.

[11] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018.

[12] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[14] Kfir Eliaz, Ran Spiegler, and Yair Weiss. Cheating with models. *American Economic Review: Insights*, 2020.

[15] Kfir Eliaz, Ran Spiegler, and Heidi C Thysen. Strategic interpretations. *Journal of Economic Theory*, 192: 105192, 2021.

[16] Ignacio Esponda and Demian Pouzo. Berk–nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3):1093–1130, 2016.

[17] Erik Eyster and Matthew Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, 2005.

[18] Marc Faddoul, Guillaume Chaslot, and Hany Farid. A longitudinal analysis of youtube's promotion of conspiracy videos. *arXiv preprint arXiv:2003.03318*, 2020.

[19] Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

[20] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[22] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

[23] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 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*, pages 624–635, 2021.

[24] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[25] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[26] Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I Jordan. Do offline metrics predict online performance in recommender systems? *arXiv preprint arXiv:2011.07931*, 2020.

[27] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[28] Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI*, 2016.

[29] Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th international conference on World Wide Web*, pages 951–961, 2016.

[30] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10869–10879, 2018.

[31] Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE, 2012.

[32] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, 2019.

[33] Martin Mladenov, Chih-Wei Hsu, Vihan Jain, Eugene Ie, Christopher Colby, Nicolas Mayoraz, Hubert Pham, Dustin Tran, Ivan Vendrov, and Craig Boutilier. Recsim ng: Toward principled uncertainty modeling for recommender ecosystems. *arXiv preprint arXiv:2103.08057*, 2021.

[34] Judea Pearl. *Causality*. Cambridge university press, 2009.

[35] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[36] Ori Plonsky, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua C Peterson, Daniel Reichman, Thomas L Griffiths, Stuart J Russell, Evan C Carter, et al. Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*, 2019.

[37] Maya Raifer, Guy Rotman, Reut Apel, Moshe Tennenholtz, and Roi Reichart. Designing an automatic agent for repeated language–based persuasion games. *Transactions of the Association for Computational Linguistics*, 10:307–324, 2022.

[38] Sven Schmit and Carlos Riquelme. Human interaction with recommendation systems. In *International Conference on Artificial Intelligence and Statistics*, pages 862–870. PMLR, 2018.

[39] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

[40] Amit Sharma, Jake M Hofman, and Duncan J Watts. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 453–470, 2015.

[41] Naeem Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012.

[42] Steven Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2005.

[43] Ran Spiegler. Behavioral implications of causal misperceptions. *Annual Review of Economics*, 12:81–106, 2020.

[44] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[45] Jessica Su, Aneesh Sharma, and Sharad Goel. The effect of recommendations on network structure. In *Proceedings of the 25th international conference on World Wide Web*, pages 1157–1167, 2016.

[46] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

[47] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[48] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[49] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34, 2021.

[50] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[51] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1717–1725, 2021.

[52] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581*, 2018.

[53] Thorsten Wuest, Daniel Weimer, Christopher Irgens, and Klaus-Dieter Thoben. Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, 4(1):23–45, 2016.

[54] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–20, 2021.

A. Additional Experiments and Results

In the main paper we present two experiments targeting user classes (*causal* or *anti-causal*) using real data. Here, we present an additional experiment targeting user subclasses (*believers* and *skeptics*) using synthetic data. Appendix D includes further details on model architectures, training procedures, and data generation.

A.1. Learning with multiple user subclasses

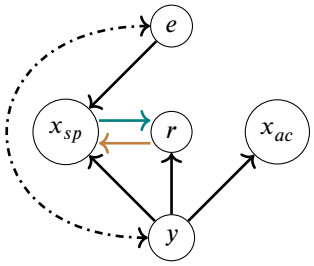

Figure 5: Data-generating process for the user subclass experiment (synthetic). Here, x factorizes into an *anti-causal* component x_{ac}, and a spurious component x_{sp} linked with r. Spuriousness results from selection bias between y and e. The system doesn't know *ex-ante* how to separate x_{ac}, x_{sp}.

Table 1: Accuracy for the user subclass experiment. Rows show train conditions: with and without regularization, and which users are included in the training set. Columns show test conditions: ID/OOD, and user type. Best results for each train condition (rows) are highlighted in bold.

Reg.	Users@train	Accuracy (ID / OOD) *skeptic*	Accuracy (ID / OOD) *believer*
$\lambda = 0$	*skeptic*	**78.0** / 50.0	**89.8** / 75.1
	believer	**78.0** / 50.0	**89.8** / 75.1
	both	**78.0** / 50.0	**89.8** / 75.1
$\lambda > 0$	*skeptic*	71.1 / **75.6**	74.67 / 75.5
	believer	69.8 / 52.5	88.5 / **85.2**
	both	70.03 / 64.57	78.88 / 78.13

Our final experiment studies learning with users of of the same-class (here, *anti-causal*) but different sub-classes: *skeptics* or *believers*. Our analysis in §3.2 suggests that each user sub-class may have a different optimal predictor; here we investigate this empirically on synthetic data.

Data. The data-generating process is as follows (see Fig. **??**). We use three environments: e_1, e_2 at train, and e_3 at test, and implement a selection mechanism (dashed line) that causes differences in $p(y|e_i)$ across e_i. Since we focus on anti-causal users, features x, r are determined by e, y. We use three binary features: x_{sp} ('spurious'), x_{ac} ('anti-causal'), and r. These are designed so that an f which

uses x_{ac} alone obtains 0.75 accuracy, but using also x_{sp} improves *in-distribution* (ID) accuracy slightly to 0.78, and so the optimal ID predictor for both user subclasses is of the form $f^*(x_{ac}, x_{sp})$. However, relying on x_{sp} causes *out-of-distribution* (OOD) performance to deteriorate considerably; thus, robust models should not learn to discard x_{sp}. The role of r is to distinguish between user subclasses: The skeptic does not need r since, for her, it is fully determined by x_{sp}; the optimal invariant predictor is hence $f_{x\to r}(x_{ac})$. Meanwhile, the believer, due to the v-structure $r\to x_{sp}\leftarrow e$, can benefit in-distribution by using both r and x_{sp}; here, the optimal invariant predictor is $f_{x\leftarrow r}(x_{ac}, r)$.

Results. Table 1 shows ID and OOD performance for each user subclass (columns), for learning with and without regularization (rows). Since all users are anti-causal, we use anti-causal (i.e., conditional) regularization. We compare learning a separate predictor for each user type (rows '*skeptic*' and '*believer*') and learning a single predictor over all users jointly ('both'). Results show that without regularization ($\lambda = 0$), ID performance is good, but the learned predictor fails OOD—drastically for skeptic users (note all rows are the same since both user types share the same ID-optimal f^*). In contrast, when regularization is applied ($\lambda > 0$), learning an independent predictor for each user subclass performs well OOD (for both subclasses), indicating robustness to changing environments; note that ID performance is also mostly maintained. Meanwhile, learning on the entire dataset (i.e., including both user types) does provide some robustness—but is suboptimal both ID and OOD.

B. Details on Formal Claims

Our claim in Proposition 1 is also based on the setting of Veitch et al. [49]. Under the assumption that e is discrete, Lemma 3.1 of [49] ensures that there exists a random variable $(x,r)^{\perp}_{e}$ such that $f_u(x,r)$ is CI if and only if it is $(x,r)^{\perp}_{e}$-measurable. Then we will assume that x, r can be decomposed into parts $x, r_{y\wedge e}, x, r^{\perp}_{y}, x, r^{\perp}_{e}$. Note that we do not assume that we know how to decompose our features in this manner, nor we assume anything about the semantic meaning of these components. We only assume that this decomposition exists, and then the main assumption made in [49] is that the graph in Fig. 1a conforms to the structures in Figure 6 for each user type.

We are now ready to state Proposition 1 in a more precise manner

Proposition 3. *Let f be a CI model and assume y and e are confounded (i.e. they are connected by an unobserved common cause c or by a directed path). Further assume that $D^e(x, r, y \mid u)$ is entailed by the causal models in Fig. 6 for $u = u_{\to y}$ and $u = u_{\leftarrow y}$. Then the following holds:*

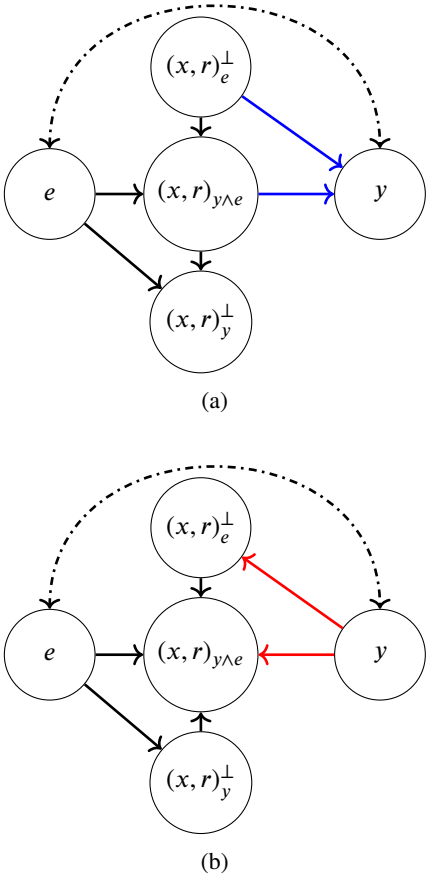

Figure 6: Detailed graphs describing our assumptions on causal and anti-causal users (a) causal model for data generating process of *causal* user, and (b) *anti-causal* user. Dashed lines indicate possible confounding.

1. $f_{u_{\to y}}$ *must satisfy* $D^e(f_{u_{\to y}}(x,r)) = D^{e'}(f_{u_{\to y}}(x,r))\ \forall e, e' \in \mathcal{E}$.
2. $f_{u_{\leftarrow y}}$ *must satisfy* $D^e(f_{u_{\leftarrow y}}(x,r) \mid y) = D^{e'}(f_{u_{\leftarrow y}}(x,r) \mid y)\ \forall e, e' \in \mathcal{E},\ y \in \{0,1\}$.

On the other hand, $f_{u_{\leftarrow y}}$ and $f_{u_{\to y}}$ do not necessarily satisfy conditions 1 and 2, respectively.

Proof. Under the assumptions laid out about the causal model, the conditional independence relations can be read off the graph directly, as in Theorem 3.2 of [49]. This proves that the independence properties stated in the proposition must hold. To see that $f_{u_{\leftarrow y}}, f_{u_{\to y}}$ do not necessarily satisfy properties 1 and 2 respectively, we will prove the existence of such cases. Consider a causal model where e and y are confounded, and assume that the model is faithful [34] (i.e. all conditional independence statements that are not entailed by the graph do not hold). Hence for the causal user we

generally have $D^e((x,r)^{\perp}_e \mid y, u = u_{\rightarrow y}) \neq D^{e'}((x,r)^{\perp}_e \mid y, u = u_{\rightarrow y})$ (at the very least there are values of $(x,r)^{\perp}_e, y$ for which this holds), and hence there exists some $(x,r)^{\perp}_e$-measurable function $\hat{f}_{u_{\rightarrow y}}(x,r)$ that satisfies $D^e(f(\hat{x},r) \mid y, u = u_{\rightarrow y}) \neq D^{e'}(f(\hat{x},r) \mid y, u = u_{\rightarrow y})$. The same argument can be applied for the anti-causal user $u_{\leftarrow y}$ to prove the existence of an $(x,r)^{\perp}_e$-measurable function $\hat{f}_{u_{\leftarrow y}}(x,r)$ that satisfies $D^e(f(\hat{x},r) \mid u = u_{\leftarrow y}) \neq D^{e'}(f(\hat{x},r) \mid u = u_{\leftarrow y})$. The model $\hat{f}(x,r)$ is CI since the constructed functions are $(x,r)^{\perp}_e$-measurable, but models $f_{u_{\leftarrow y}}, f_{u_{\rightarrow y}}$ do not satisfy conditions 1 and 2 respectively, which concludes our claim. □

Next we prove Proposition 2 by constructing a confounded model for an anti-causal user, similar to the one in the synthetic experiment of Section A.1. Towards this proposition, we point out that an optimal CI predictor is defined as a CI predictor with the best possible worst case performance. Where the worst case is taken over all distributions that are causally-compatible [49] with the source distribution D_{train}.

Definition 1. *D_{train} and D_{OOD} are causally compatible if they are entailed by the same causal graph, $D_{train}(y) = D_{OOD}(y)$, and there is a confounder c and/or selection conditions $s, \tilde{s}$ such that $D_{train} = \int D_{train}(x_{sp}, x_{ac}, r, y \mid c, s = 1) d\tilde{P}(c)$ and $D_{OOD} = \int D_{train}(x_{sp}, x_{ac}, r, y \mid c, \tilde{s} = 1) d\tilde{Q}(c)$ for some $\tilde{P}(c), \tilde{Q}(c)$.*

Let us focus now on distributions where $f(x_{sp}, r, x_{ac})$ is counterfactually invariant if and only if it is (r, x_{ac})-measurable (the expression (r, x_{ac}) should be read as a bivariate random variable). Note again that from Lemma 3.1 of [49] such a variable exists. The following claim will help us reason about the optimal CI model for users of the skeptic sub-class.

Lemma 1. *If D_{train} is entailed by the graph in Fig. 7a and D_{OOD} is causally compatible with it, then $D_{train}(y \mid r, x_{ac}) = D_{OOD}(y \mid r, x_{ac})$.*

Proof. For binary classification, it is enough to show that $\frac{D_{\text{train}}(y=1|r,x_{ac})}{D_{\text{train}}(y=0|r,x_{ac})} = \frac{D_{\text{OOD}}(y=1|r,x_{ac})}{D_{\text{OOD}}(y=0|r,x_{ac})}$. Let us write this for the training distribution:

$$\begin{aligned}\frac{D_{\text{train}}(y=1 \mid r, x_{ac})}{D_{\text{train}}(y=0 \mid r, x_{ac})} &= \frac{D_{\text{train}}(r, x_{ac} \mid y=1) D_{\text{train}}(y=1)}{D_{\text{train}}(r, x_{ac} \mid y=0) D_{\text{train}}(y=0)} \\ &= \frac{D_{\text{train}}(r, x_{ac} \mid y=1) D_{\text{OOD}}(y=1)}{D_{\text{train}}(r, x_{ac} \mid y=0) D_{\text{OOD}}(y=0)}.\end{aligned}$$

The second equality stems from the causal-compatibility of D_{OOD}. It is left to show that $D_{\text{train}}(y \mid r, x_{ac}) = D_{\text{OOD}}(y \mid r, x_{ac})$. From causal-compatibility the distributions are entailed by the same graph in Fig. 7a, which imposes the conditional independence $c \perp r, x_{ac} \mid y$. Hence we conclude the proof by:

$$\begin{aligned}D_{\text{train}}(x_{ac}, r \mid y) &= \int D_{\text{train}}(x_{ac}, r \mid y, c) d\tilde{P}(c) \\ &= \int D_{\text{train}}(x_{ac}, r \mid y, c) d\tilde{Q}(c) \\ &= D_{\text{OOD}}(x_{ac}, r \mid y).\end{aligned}$$

□

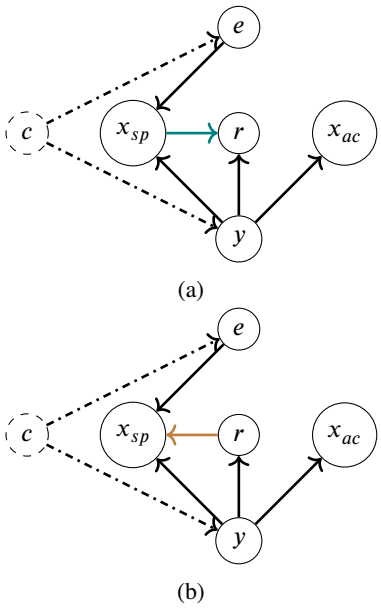

Figure 7: Graphs describing the data-generating processes for anti-causal *believer* and *skeptic* users in the proof of Proposition 2.

From this result we gather that if we only consider the features x_{ac}, r, there is a unique Bayes-optimal classifier over all target distributions that are causally compatible with D_{train}. Since a classifier is CI if and only if it is (x_{ac}, r)-measurable, we see that for the skeptic sub-class of users the optimal CI model is $f(x_{sp}, r, x_{ac}) = D_{\text{train}}(y \mid r, x_{ac})$. The rest of the proof will simply show that this model may not be CI for a user of sub-type *believer* that has the same choice patterns over observed data pooled from two training environments.

Proof of Proposition 2. Consider a data generating process as depicted in Figure 7a. All variables x_{sp}, r, x_{ac}, y, c are binary, we consider 2 training environments $\mathcal{E}_{\text{train}} = \{0, 1\}$.

We write down the distribution in a factorized form:

$$D_{u_{x\leftarrow r}}(x_{sp}, x_{ac}, r, y) = \sum_{c\in\{0,1\}, e\in\{0,1\}} p(c)p(y \mid c)p(x_{ac} \mid y)p(e \mid c)p_{u_{x\leftarrow r}}(r \mid y)p^e_{u_{x\leftarrow r}}(x_{sp} \mid r, y)$$

$$= p(x_{ac} \mid y)p_{u_{x\leftarrow r}}(r \mid y)\left(\sum_{e\in\{0,1\}} \tilde{p}(e, y)p^e_{u_{x\leftarrow r}}(x_{sp} \mid r, y)\right).$$

Here we defined $\tilde{p}(e, y) = \sum_{c\in 0,1} p(y, c)p(e \mid c)$. The subscripts $u_{x\leftarrow r}$ emphasize that in the distribution we will construct for the believer user, $D_{u_{x\to r}}$, all factors that are not subscripted will be equal to those in $D_{u_{x\leftarrow r}}$. That is, consider a distribution that factorizes over the graph in Figure 7b as follows:

$$D_{u_{x\to r}}(x_{sp}, x_{ac}, r, y) = \tag{1}$$
$$p(x_{ac} \mid y)p_{u_{x\to r}}(r \mid y, x_{sp})\left(\sum_{e\in\{0,1\}} \tilde{p}(e, y)p^e_{u_{x\to r}}(x_{sp} \mid y)\right).$$

We will show that there exists some setting of $p_{u_{x\to r}}(r \mid y, x_{ac}), p^e_{u_{x\to r}}(x_{sp} \mid y)$ such that:

$$D_{u_{x\leftarrow r}}(x_{sp}, x_{ac}, r, y) = D_{u_{x\to r}}(x_{sp}, x_{ac}, r, y).$$

But it will also satisfy $D^0_{u_{x\to r}}(y \mid r, x_{ac}) \neq D^1_{u_{x\to r}}(y \mid r, x_{ac})$. Then the proof will be concluded, as $f(x_{sp}, x_{ac}, r) = D_{u_{x\leftarrow r}}(y \mid r, x_{ac}) = D_{u_{x\to r}}(y \mid r, x_{ac})$ cannot be CI w.r.t $D_{u_{x\to r}}$. This holds since $D^e_{u_{x\to r}}(y \mid r, x_{ac}) \neq D_{u_{x\to r}}(y \mid r, x_{ac})$ for $e \in \{0, 1\}$, hence there must be some instance for which $f(x_{ac}(0), x_{sp}(0), r(0)) \neq f(x_{ac}(1), x_{sp}(1), r(1))$.

Towards this, consider $D_{u_{x\leftarrow r}}(r \mid y, x_{sp})$ which is obtained by the respective marginalization and conditioning of $D_{u_{x\leftarrow r}}(x_{sp}, x_{ac}, r, y)$, and also consider $\sum_{e\in 0,1} \tilde{p}(e, y)D^e_{u_{x\leftarrow r}}(x_{sp} \mid y)$. Let us set:

$$p_{u_{x\to r}}(r \mid y, x_{sp}) := D_{u_{x\leftarrow r}}(r \mid y, x_{sp}).$$

It is clear that if we set $p^e_{u_{x\to r}}(x_{sp} \mid y)$ such that the following holds:

$$\sum_{e\in\{0,1\}} \tilde{p}(e, y)p^e_{u_{x\to r}}(x_{sp} \mid y) = \sum_{e\in\{0,1\}} \tilde{p}(e, y)D^e_{u_{x\leftarrow r}}(x_{sp} \mid y), \tag{2}$$

then the equality $D_{u_{x\leftarrow r}}(x_{sp}, x_{ac}, r, y) = D_{u_{x\to r}}(x_{sp}, x_{ac}, r, y)$ also holds. That is because the factorization in (1) is a factorization of the joint distribution over x_{sp}, x_{ac}, r, y where all factors are equal to the ones obtained from $D_{u_{x\leftarrow r}}(x_{sp}, x_{ac}, r, y)$. [1]

[1] Note that it is easy to observe that the two sides of (2) are the marginal distribution over x_{sp}, y of the two distributions $D_{u_{x\to r}}$ and $D_{u_{x\leftarrow r}}$ respectively.

Finally, we claim that many solutions satisfy (2). For each value of y, x_{sp} Eq. (2) is a linear equation with two variables ($p^0_{u_{x\to r}}(x_{sp} \mid y)$ and $p^1_{u_{x\to r}}(x_{sp} \mid y)$), and they should be constrained to take values in the range $[0, 1]$. One solution to the equation is to set $p^e_{u_{x\to r}}(x_{sp} \mid y) := D^e_{u_{x\leftarrow r}}(x_{sp} \mid y)$, and unless $D^e_{u_{x\leftarrow r}}(x_{sp} \mid y) \in \{0, 1\}$ for each value of x_{sp}, y, and $D^0_{u_{x\leftarrow r}}(x_{sp} \mid y) = D^1_{u_{x\leftarrow r}}(x_{sp} \mid y)$ (i.e. the spurious feature completely determines y) the set of solutions to the equations forms an interval in $\mathbb{R}^2$, and has Lebesgue measure that is non-zero.

Thus let us consider the set of parameterized (by the factors in (1)) distributions $\tilde{D}_{u_{x\to r}}(e, x_{sp}, r, x_{ac}, y)$ that satisfy $\sum_{\tilde{e}} \tilde{D}_{u_{x\to r}}(e = \tilde{e}, x_{sp}, r, x_{ac}, y) = D_{u_{x\leftarrow r}}(x_{sp}, r, x_{ac}, y)$ for the fixed distribution $D_{u_{x\leftarrow r}}(x_{sp}, r, x_{ac}, y)$. This set has a non-zero Lebesgue measure over the linearly independent parameters needed to parameterize $D_{u_{x\to r}}$. Since the set of parameters that yield unfaithful distributions w.r.t a graph has Lebesgue measure zero [44], there must be at least one distribution $\tilde{D}_{u_{x\to r}}(e, x_{sp}, r, x_{ac}, y)$ in the set where the independence $r, x_{ac} \perp e \mid y$ does *not* hold. For such a distribution we will have $D^e_{u_{x\to r}}(y \mid r, x_{ac}) \neq D_{u_{x\to r}}(y \mid r, x_{ac})$, which is what was required to conclude the proof.

□

C. Related Work

Causality and Recommendations. Formal causal inference techniques have been used extensively in many domains, but have only recently been applied to recommendations [28,52,7,54,51]. Liang et al. [29] use causal analysis to describe a model of user exposure to items. Some work has also been done to understand the causal impact of these systems on behavior by finding natural experiments in observational data [40,45,39], and through simulations [11,38]. Bottou et al. [8] use causally-motivated techniques in the design of deployed learning systems for ad placement to avoid confounding. As most of this literature addresses selection bias and the effect of recommendations on user behavior [7,54,51], there is no work, as far as we know, that models boundedly rational agents interacting with a recommender system. Moreover, we are the first to propose modeling users' (mis)perceptions about the recommendation generation process using causal graphs.

Bounded Rationality and Subjective Beliefs. The bounded rationality literature focuses on modelling agents that make decisions under uncertainty, without the ability to fully process the state of the world, and therefore hold subjective beliefs about the data-generating process. Eyster and Rabin [17] defined *cursed beliefs*, which capture an agent's failure to realize that his opponents' behavior depends on factors beyond those he is informed of. Building on Esponda

Table 2: Original *RateBeer* dataset statistics.

Number of reviews	2,924,127
Number of users	40,213
Number of beers	110,419
Users with > 50 reviews	4,798
Median #words per review	54
Timespan	4/2000-11/2011

and Pouzo [16], who modelled equilibrium beliefs under misspecified subjective models, Spiegler [43] used causal graphs to analyze agents that impose subjective causal interpretations on observed correlations. This work lays the foundation upon which we model users here, and has sprouted many interesting extensions [14,15].

Causality and Invariant Learning. Correlational predictive models can be untrustworthy [23], and latch onto spurious correlations, leading to errors in OOD settings [32,20,19]. This shortcoming can potentially be addressed by a causal perspective, as knowledge of the causal relationship between observations and labels can be used to mitigate predictor reliance on them [9,49]. In our experiments, we learn a representation that is invariant to interventions on the 'environment' e, a special case of an invariant representation [3,27,4]. Learning models which generalize OOD is a fruitful area of research with many recent developments [30,22,35,46,5,50]. Recently, Veitch et al. [49] showed that the means and implications of invariant learning depend on the data's true causal structure. Specifically, distinct causal structures require distinct regularization schemes to induce invariance.

D. Experimental Details

Code and data for all experiments can be found in the following anonymous link:
https://drive.google.com/drive/folders/1bO57v4PUuUh76F_q0a_xAVx6CKdeDJ5l

D.1. *RecBeer* (causal users)

Original Dataset description. The original *RateBeer* dataset includes textual reviews and numerical ratings of roughly 3000 unique beers, collected over the span of over 11 years. Each review data-point also includes additional features describing the beer (e.g., brand, style), the author of the review (e.g., location), and the review itself (e.g., date). Figure 8 shows an example of a data point. Table 2 provides summary statistics.

Data Generation Process. The original RateBeer dataset includes reviews and rating that were authored and submitted by users of the platform. For our purposes, focusing learning and prediction on users as *contributors* of content has two limitations: (i) we cannot know what platform-selected infor-

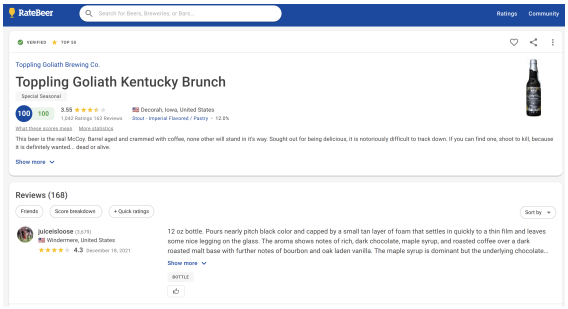

Figure 8: ***RateBeer* example:** A textual review and numerical rating for a beer (with metadata).

Figure 9: ***RecBeer* interventions:** An example of a simulated intervention for causal users, for which changing the review shown to the user (bottom) to another (top) may influence his behavior (here, from not choosing to choosing).

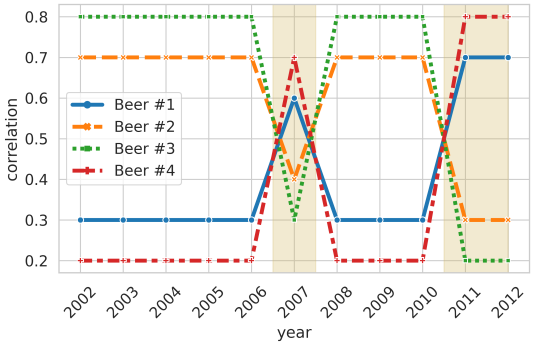

Figure 10: ***RecBeer* environments:** Each year serves as a different environment, whose affect is expressed through differing correlations between beer types and user choices. The plot shows the temporal correlation structure used for the experiment in §4.1, and underlie the results presented in Fig. 3. Periods with substantial changes are highlighted in tan.

Table 3: Our *RecBeer* data features.

Variable type	Not.	Description
Item	x	avg past appearance avg past aroma avg past palate avg past taste # of active years alcohol percentage beer type
User	u	avg past satisfaction # of past choices # of active years
Recommendation	r	text review # of past reviews
Time	e	year
Choice	y	try beer/not

mation (r) was presented to them and how it influenced their decisions, and (ii) we cannot reason counterfactually about their potential choices had they been exposed to different information.

To overcome both issues, we adapt the original dataset to simulate choice behavior of users as *consumers* of content, as they use the platform to make informed decisions about beer consumption. We emulate the following process: a user u logs on to the platforms, and is recommended a certain beer. The beer is described by intrinsic features x, and one platform-selected textual review r, chosen from a pool of already-existing reviews for that beer (these being the reviews for that beer that have already by submitted by other contributing users). The user then decides weather to try (i.e., consume) the beer ($y = 1$) or not ($y = 0$). Our goal is to predict for new users u their choices y for recommended beers given descriptions x, r.

To create features for beers x and (consuming) users u, we aggregate information from all corresponding reviews: for beers—all reviews of that beer, and for users—all reviews authored by that user. This includes features such as average past taste score for beers and average past overall satisfaction for users. Table 3 summarizes our feature space. Since we model users as *causal*, the graph edge $r \to y$ implies that changes to r causally affect y. To simulate this behavior, we create for each user an 'intervention space' which includes a collection of possible interventions r and their corresponding counterfactual outcomes y. For our experiment, we simply take all pairs of reviews and ratings (r, s) for a given beer to be the set of possible interventions and outcomes. Textual reviews are featurized using a pre-trained BERT model [13], and numerical ratings $s \in [0, 5]$ are transformed into binary choices $y = \{0, 1\}$ by setting $y = 1$ if the user's rating for that beer was above the median rating (for that beer), and $y = 0$ otherwise. Since learning requires observational data, for each user-beer pair (u, x) we sample (in a way we describe shortly) one review-choice pair (r, y) out of 100 unique reviews for that beer; an example is presented in Figure 9. This provides a sampled tuple (u, x, r, y) expressing the behavior of a *causal* user whose choices are affected by the review presented to her. Together, u, x, and r (as an embedding) include 866 features.

Finally, to model the effects of changing environments, we consider an environment variable e that encodes the year, expressing the idea that different years may express different 'trends' in which beer *types*[2] are more (and less) fashionable. To implement this, we sample review-choice pairs for users within each year in a way that introduces a pre-determined amount of correlation between choices and beer types. The chosen per-year correlation levels is plotted in Figure 10. Notice the drastic change in fashions in 2007 and 2011.

Training and testing. We train and evaluate one model per year. For each year $e \in \{2006, \ldots, 2012\}$, training is performed on data from years $\{2002, \ldots, e-1\}$ and tested on e. In this way, fashions regarding beer type accumulate over time.

Models. We learn a linear model that takes as input the concatenation of u, x, r. The learning objective includes a binary cross entropy loss, and marginal MMD as regularization [21] (since we model users as causal; see §3). We trained all models for 700 epochs with $lr = 0.01$ and batches of size 1024, and set $\lambda = 100$. Results are averaged over five runs with different random seeds.

D.2. *RecFashion* (anti-causal users)

Original Dataset Statistics. The *Fashion Product Images* dataset includes a large collection of fashion items, described by an image and additional attributes such as: season, gender, base color, usage, year, and product display name. Items are organized by category, sub-category, and type; we focus on the *apparel* category. Table 4 provides summary statistics.

Table 4: Original *Fashion Product Images* dataset statistics.

number of items	44,447
main categories	7
sub-categories	45
types	142

[2]We create four beer 'types' by aggregating beers of similar style. For example, the styles *Doppelbock*, *Dortmunder*, *Dunkel*, *Dunkelweizen*, and *Dunkler* were all attributed to the same type.

Figure 11: Fashion items in the *RecFashion* dataset with recommended colors. On the left side are green recommendations and on the right side are red recommendations.

Data Generation Process. The original dataset does not include user choices (or any other form of user behavior). To simulate user choices, we imagine a setting were the platform recommends to each user an item by presenting an image of the item (x) in a certain color (r). We set x to be the item's grayscale image, and set r to be a colorization of that image into one of two colors: red or green. Users then choose whether to buy the item or not, $y \in \{0, 1\}$. We then model users as choosing primarily on the basis of the 'gender' attribute of items, $x_g \in \{0, 1\}$, and set $y = x_g$ w.p. 0.75 and $y = 1 - x_g$ otherwise.

Since users in this experiments are anti-causal, they act under the belief that changes in y affect r (here we do not make use of the edge $y \to x$). Note that e also affects r. We implement this joint influence of e, y on r by assigning colors to images in a way that obtains a certain level of correlation between the color $r \in \{\text{red}, \text{green}\}$ and choices y. Technically, we associate with each environment e a parameter $p_e \in [0, 1]$. Then, using a color variable $c = 0$ for red and $c = 1$ for green we assign for each item its color as $c = y$ w.p. p_e, and $c = 1 - y$ otherwise. Thus, different environments entail different conditional distributions $P(r = \text{red}|y = 1) = P(r = \text{green}|y = 0) = p$, which reflect an anti-causal structure. Finally, given the sampled c, we colorize the image x as follows: if $c = 1$, we set $x_R \leftarrow 0.5{+}0.2x_R$, $x_G \leftarrow 0.7x_G$, $x_B \leftarrow 0.7x_B$; if $c = 0$, we set $x_G \leftarrow 0.5 + 0.2x_G$, $x_R \leftarrow 0.7x_R$, $x_B \leftarrow 0.7x_B$ (R, G, B are the color channels). Note that this means users do not observe x, r independently, but rather a colored image that is a product of both x and r.

Training and testing. We run eight experiments that differ in the average degree of correlation in the training sets, for average correlation values of $p \in \{0.1, 0.2, \ldots, 0.8\}$. Each experimental condition (p) includes training data from six environments e, with correlations $p_e nv \in \{p - 0.025, p + 0.025, p - 0.05, p + 0.05, p - 0.1, p + 0.1\}$ (their average is p).

Models. For the model We used a feed forward neural network with three hidden layers and a hidden dimension of size 256, ReLU activation function and NLL as our base loss function. For computational efficiency, input images were resized to 14×14. The learning objective includes a binary cross entropy loss, and a conditional DeepCORAL regularizer [47] (since we model users as anti-causal; see §3). We set $\lambda = 5000$ in the first 125 epochs and $\lambda = 1$ in the rest, and trained the model for 1,900 epochs with $lr = 0.001$ and batches of size 1024.

E. Loss Functions.

We train all of our models with either the *CORAL* or *MMD* loss. Empirically, we found that *CORAL* we more stable in the *RecFashion* experiments and. In the *RecBeer* experiments, models trained with the *MMD* loss consistently outperformed those who were not. When conditioning on the label y, we compute l_{dist} (either l_{CORAL} or l_{MMD}) separately for cases where $y = 1$ and $y = 0$. We describe here both loss functions.

***CORAL* Loss.** The *CORAL* loss is the distance between the second-order statistics of two feature representations, corresponding to different z:

$$l_{CORAL}(f(x, r), z) = \frac{1}{d^2}||C_z - C_{z'}||_F^2$$

where $||\cdot||_F^2$ denotes the squared matrix Frobenius norm. The covariance matrices of the source and target data are given by:

$$C_z = \frac{1}{n_z - 1}(\phi(x(z), r)^\top \phi(X(z), r) - \frac{1}{n_z}(\mathbf{1}^\top \phi(x(z), r))^\top (\mathbf{1}^\top \phi(x(z), r)))$$

where $\mathbf{1}$ is a column vector with all elements equal to 1, and $\phi(\cdot)$ is the feature representation.

***MMD*.** Maximum mean discrepancy (*MMD*) measures distances between mean embeddings of features. That is, when we have distributions P and Q over a set $\mathcal{X}$. The *MMD* is defined by a feature map $\phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is what's called a reproducing kernel Hilbert space. In general, the *MMD* is

$$\text{MMD}(P, Q) = ||\mathbb{E}_X[\phi(X)] - \mathbb{E}_Y[\phi(Y)]||_{\mathcal{H}}$$

For use of the MMD loss for causal representation learning, see Veitch et al. [49].