Pragmatic Reasoning in Design

Lance Ying[1,2], **William Van Uitert**[1], **Tan Zhi-Xuan**[3],
Joshua B. Tenenbaum[2], **Samuel J. Gershman**[1]

[1]Harvard University, [2]Massachusetts Institute of Technology, [3]National University of Singapore

Abstract

People can often understand and use novel artifacts after only a few interactions, suggesting that design choices communicate underlying affordances and causal structure. We propose a formal account of this process by framing cooperative, user-centered design as a cooperative game in which the user is the principal and the designer is an assistant. Inspired by prior work on pragmatic communication (e.g. RSA), our model treats a designer's design decisions as communicative signals and predicts user judgments via recursive mentalizing: designers make design decisions to trade off informativeness about the artifact with efficiency, and users infer the true model of the artifact by inverting this cooperative designer model. We evaluate the model in a "design game" where designers place visually identical keys on trays to help a user infer which keys unlock which doors in grid-world layouts. We find that pragmatic designer and user models better match human judgments than non-mentalizing literal baselines.

Keywords: Theory of Mind, design, pragmatic communication, assistance game

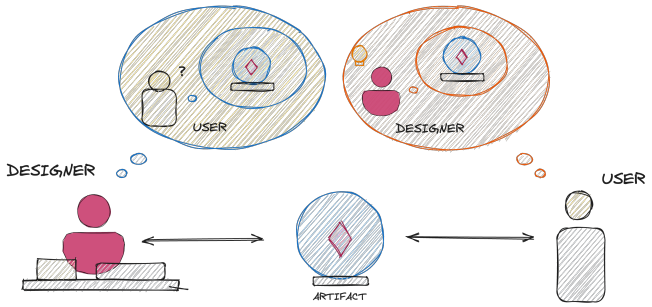


Figure 1: Illustration of recursive mentalizing in user-centered design, where the designer optimizes the artifact by simulating how the user may interpret the design, and the user leverage this process to infer the underlying model of the artifact by assuming it is the product of rational and intentional decision making by a cooperative mentalizing designer.

Introduction

People routinely learn to use unfamiliar artifacts after only a few interactions, suggesting that design choices provide cues about an artifact's affordances and underlying causal structure. A central challenge for cognitive science and human-computer interaction (HCI) research is to explain how users form these mental models, and how designers can intentionally shape them.

Prior work in user-centered design emphasizes that designers should anticipate users' goals, limitations, and likely confusions when creating an artifact (Norman, 2013). Despite interests in the intersection between design and cognitive science, there has not been a formal account of this process, and it's unclear how users interpret and interact with artifacts that are intentionally designed for them.

While previous computational accounts of pragmatic communication have focused on *communicative intent recognition*—inferring an agent's hidden communicative intent from observed actions, many design settings differ in a crucial way: goals are often *common knowledge*. The designer knows the user's goal and aims to help the user achieve it, and the user expects the designer to be helpful (see Figure 2 for an example).

This shifts the problem from "What does the other agent want?" to "How does this artifact work?." Design often communicates a *mental model* of an artifact—what actions are possible and what outcomes they produce—rather than a private intention.

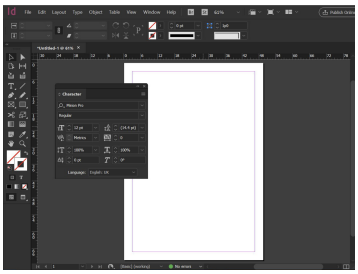
We propose a formal model that frames cooperative design as a *design game* in which the *user* is the principal and the *designer* is an assistant. Inspired by pragmatic communication models such as the Rational Speech Act (RSA) framework (Goodman & Frank, 2016), we treat design choices as communicative actions: the designer recursively mentalizes about how the user will interpret available cues, and the user reasons about what a helpful designer would have built (Figure 1).

We instantiate this framework in a room design game across a range of rooms and scenarios, where a knowledgeable designer places visually identical keys on trays in a partially filled grid-world rooms to help a naive user infer which keys unlock which doors. We develop pragmatic designer and pragmatic user models, along with non-mentalizing literal baselines.

We find that the pragmatic designer model predicts human design choices substantially better than the literal baseline. Together, these results support the claim that designers can reason about the users' goals and actions when designing an environment, while users can interpret novel designs by reasoning not only about perceptual cues, but also about a helpful designer's communicative and cooperative intent.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

(a) AirBnB key placement



(b) Adobe InDesign user interface.

(c) Room light switches

Figure 2: Example designs and pragmatic inference. **(a)** An AirBnB host intentionally left a key on a table with an open book. A pragmatic user can infer that the key must be related to the guest's booking. **(b)** Adobe InDesign UI which shows many buttons crowded in a toolbar. Each icon individually may offer limited information of their functionality, but by reasoning about how the icons are positioned, one may pragmatically infer additional information about the buttons. **(c)** Room design where switches on the wall may correspond to a particular appliance. A pragmatic reasoner may assume that the switches are designed to help one easily navigate a room.

Related Work

Our work connects research on: (i) design as communication and interpretability, (ii) pragmatic models of inference in language, and (iii) cooperative/assistance games in multi-agent interaction.

User-Centered Design A central idea in HCI is that successful artifacts are designed around users' goals, limitations, and expectations rather than around internal system structure. Classic accounts of user-centered design emphasize that designers should anticipate the user's actions and likely confusions, making the correct actions easy to discover and execute while preventing or mitigating errors (Gibson, 2014; Nielsen, 1993; Norman, 2013).

A practical implication is that designers often need to *simulate* users: they evaluate candidate designs by predicting how a user would interpret available cues and what actions would follow. Methods such as cognitive walkthrough operationalize this idea by stepping through tasks from the perspective of a novice user and asking whether the interface makes the next action sufficiently salient (Wharton et al., 1994). Related cognitive modeling approaches (e.g., GOMS) similarly aim to predict usability and learning costs by approximating users' internal procedures (Card et al., 1983).

Pragmatic Communication We build on work in *pragmatic communication* that treats communicative behavior as rational action shaped by an observer's inference. In the Rational Speech Act (RSA) framework, speakers choose utterances to be informative with respect to a listener model, and listeners invert that process to infer intended meaning (Frank & Goodman, 2012; Goodman & Frank, 2016). A closely related tradition models social inference as *inverse planning*: observers infer hidden goals and intentions by assuming agents act (approximately) rationally to achieve them (Baker et al., 2009; Zhi-Xuan et al., 2020). Recent work extends this idea to settings where agents intentionally shape others' inferences ("inverse inverse planning"; Chandra et al., 2023).

Work on *pedagogy* also frames teaching as sequential decision making: teachers select demonstrations or interventions to guide a learner's beliefs and policies, and learners interpret those demonstrations as communicative (Chen et al., 2024; Ho et al., 2018; Shafto et al., 2014). More broadly, people can infer communicative intent from non-linguistic actions and artifacts, including how objects are placed in the environment (Lopez-Brau & Jara-Ettinger, 2023; Royka et al., 2022).

Assistance Games Assistance games and cooperative-inference formulations treat interaction as a team problem with asymmetric information: one agent knows the task or dynamics and must act so that a partner can succeed (Fisac et al., 2019; Hadfield-Menell et al., 2016; Zhi-Xuan et al., 2024). This perspective has been influential in multiagent systems, human-centered AI, and AI alignment, where agents are evaluated by how well they enable a human to achieve goals rather than by their own direct reward. Within this paradigm, human users know their own goals, but (AI) assistants do not, so assistants must infer goals from user behavior, while users should legibly demonstrate their goals (Dragan & Srinivasa, 2013) in order to receive helpful assistance.

A related line of work formalizes the idea that an observer may need to infer the intended objective from an imperfect specification; for example, inverse reward design treats a provided reward function as evidence about the designer's true preferences under a model of bounded specification (Hadfield-Menell et al., 2017). In contrast to both assistance games and inverse reward design, we assume that the user's objective is known to the designer/assistant. Instead, it is the designer who possesses hidden knowledge about artifact function that the user must infer.

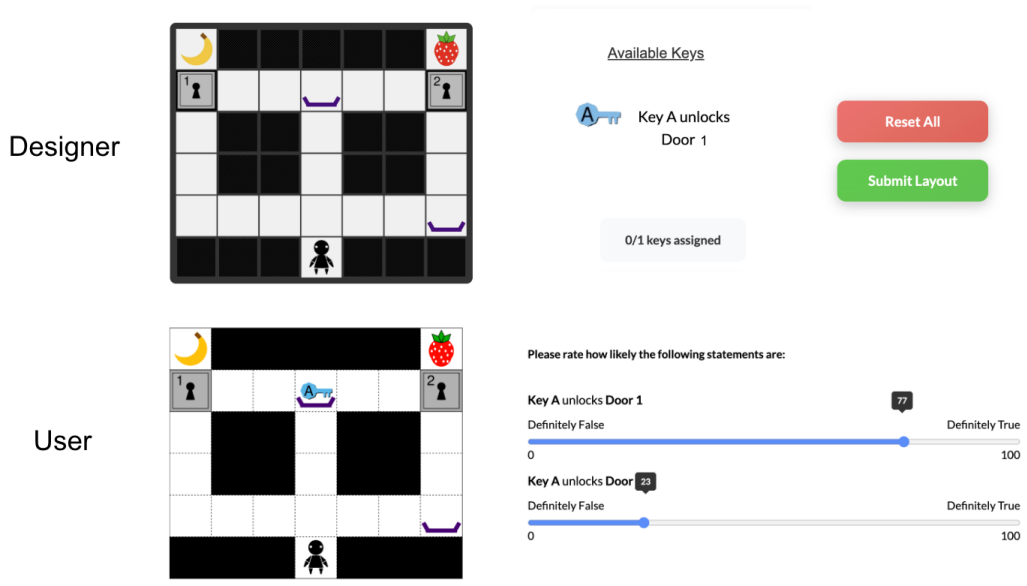


Figure 3: Experimental interface for the Room Design Game. (**Top**) *The designer phase*: Human participants are recruited to act as designers in the game. They are provided a set of key(s) and the ground truth information on which door each key unlocks. The human participants are asked to drag keys into purple key trays in order to help the player (black agent). (**Bottom**) *The user phase*: Human participants are recruited to act as player (black agent) in the game. They are told that the room was designed by another human to help them achieve the goal. The user agent is then asked to rate statements on what door they believe each key unlocks on a continuous scale.

Design as a Cooperative Game

Inspired by prior work on rational communication (Frank & Goodman, 2012) and assistance games (Hadfield-Menell et al., 2016), we formalize cooperative design as a two-player cooperative game — a *design game* — between a **designer** (D) and a **user** (U). The game takes place in an environment ε consisting of three components: an initial state s_0, commonly-known causal structure $\varepsilon_{\text{common}}$, and hidden causal structure $\varepsilon_{\text{hidden}}$. Together, $\varepsilon_{\text{common}}$ and $\varepsilon_{\text{hidden}}$ determine how the user's actions affect the environment's state. Only the designer knows the full environment, including $\varepsilon_{\text{hidden}}$, while the user knows only $\varepsilon_{\text{common}}$ and must infer the hidden structure from the designed artifact.

The game proceeds in two phases. In the *designer phase*, the designer modifies the initial state s_0 through a limited set of actions, producing a designed artifact s_1. We can think of s_0 as "raw materials" that the designer arranges into the artifact. In the *user phase*, the user observes s_1 and takes actions toward their goal g. We assume the designer knows the user's goal. The designer's task is to construct the artifact so as to best enable the user to achieve their goal, which may involve communicating information about $\varepsilon_{\text{hidden}}$ through the design of s_1. Conversely, the user may need to infer the hidden structure from the design in order to achieve g efficiently.

The Room Design Game

We instantiate this framework in the Room Design Game (Figure 3). The environment is a gridworld containing fruits locked behind doors $d_1, d_2, \ldots$, with trays $t_1, t_2, \ldots$ at various locations. Each door is unlocked by exactly one key; keys are labeled but visually identical. The commonly-known structure $\varepsilon_{\text{common}}$ includes movement rules, that keys can be picked up, and that keys must be on trays. The hidden structure $\varepsilon_{\text{hidden}}$ is the key-door mapping m^*: for each key k, which door d satisfies $\textsc{unlocks}(k, d)$. The user's goal g is to obtain any of the fruits, which requires unlocking at least one door.

The initial state s_0 is the gridworld with empty trays. In the *designer phase*, the designer places each key k in some tray t—denoted $\textsc{placed}(k, t)$—producing the artifact s_1. The designer knows m^* and knows the user lacks this information. In the *user phase*, the user observes s_1 and must infer the hidden mapping. We measure users' beliefs by having them rate statements of the form "key k unlocks door d" on a scale from 1 (definitely disagree) to 100 (definitely agree).

Computational Model

The core insight of our model is that a pragmatic designer balances two considerations: the signal a placement provides about the hidden key-door mapping, and the efficiency of the placement for the user's task. A pragmatic user, in turn, assumes a pragmatic designer and inverts this model to infer the hidden mapping. We first describe the literal designer and user as baselines, then introduce the pragmatic agents.

Literal Designer (D_1). The literal designer places each key in a nearby tray (in terms of maze distance), treating design

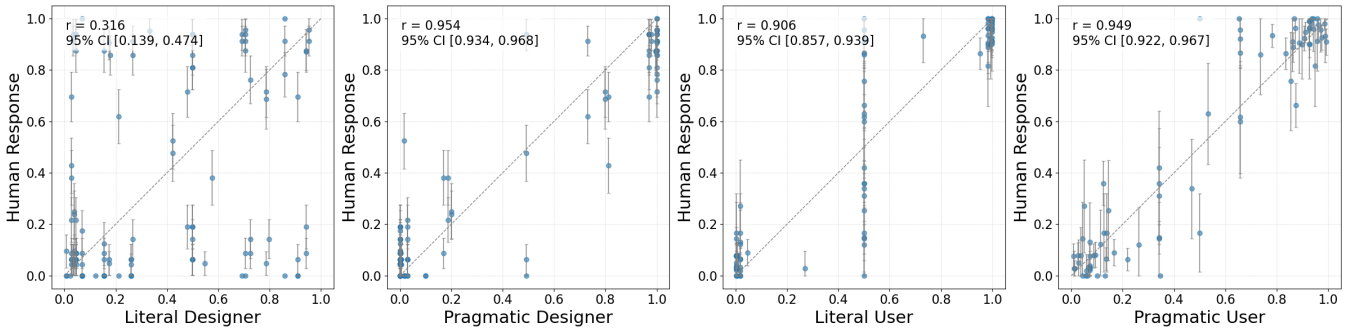


Figure 4: Model predictions compared to human judgments. Overall, we find that pragmatic models fit human judgments better than literal models in bother design and user experiments.

as efficiency optimization. Suppose $\textsc{unlocks}(k, d)$ holds in the true mapping m^*. The start position of the user is s. The utility of placing k in tray t is:

$$u_{D_1}\big(\textsc{placed}(k,t); m^*\big) = -\text{Distance}(s,t,d). \quad (1)$$

where $\text{Distance}(s,t,d)$ is the total distance of the user first going to tray t and then to unlock door d. The designer's policy is a softmax over utilities:

$$P_{D_1}\big(\textsc{placed}(k,t) \mid m^*\big) \propto \exp\big(\beta \cdot u_{D_1}\big), \quad (2)$$

where $\beta \geq 0$ is an inverse temperature parameter.

Literal User (U_1). The literal user infers which door a key unlocks based on spatial proximity alone, without considering the designer's communicative intent. Given the observation that a key k is on tray t, the literal user assumes k is more likely to unlock nearby doors rather than far ones:

$$P_{U_1}\big(\textsc{unlocks}(k,d) \mid \textsc{placed}(k,t)\big) \propto \exp\big(-\beta \cdot \text{Distance}(s,t,d)\big). \quad (3)$$

Pragmatic Designer (D_2). The pragmatic designer chooses placements that trade off informativeness — how well a placement helps the user infer the true mapping — against efficiency. Suppose key k unlocks door d in the true mapping m^*. Let $P_{U_1}(m \mid \textsc{placed}(k,t))$ be the literal user's posterior after observing placement in tray t. The designer's utility is:

$$u_{D_2}\big(\textsc{placed}(k,t); m^*\big) = -\alpha_1 \,\text{KL}\Big(P_{U_1}(m|\textsc{placed}(k,t)) \,\|\, \delta_{m^*}\Big) - \alpha_2 \,\text{Distance}(s,t,d), \quad (4)$$

where δ_{m^*} is the point mass at the true mapping, and $\alpha_1, \alpha_2 \geq 0$ weight the two terms. Their policy is:

$$P_{D_2}\big(\textsc{placed}(k,t) \mid m^*\big) \propto \exp\big(\beta \cdot u_{D_2}\big). \quad (5)$$

Pragmatic User (U_2). The pragmatic user assumes a D_2 designer and performs Bayesian inference to recover m^*:

$$P_{U_2}(m^* \mid \textsc{placed}(k,t)) \propto P_{D_2}\big(\textsc{placed}(k,t) \mid m^*\big)\, P(m^*). \quad (6)$$

That is, the pragmatic user inverts the designer's policy, reasoning about why a cooperative designer would have chosen the observed placement. We assume a uniform prior for $P(m^*)$.

Experiment Design

We designed 10 unique map layouts and generated 3 scenarios per map by varying the key-door mapping, for a total of 30 scenarios.

We recruited 60 participants for the designer study (Mean age = 40.92, 37 male, 23 female) and 50 participants for the user study (Mean age = 43.04, 15 male, 35 female). Because designs can vary across designers, we aggregate designs by taking the modal placement per trial and use that design as the stimulus for the user study. Each participant completes 10 trials in randomized order.

For computational models, we performed a grid search over all hyperparameters and retained the best fitting hyperparameters for analysis.

Results

Figure 4 compares model predictions to human responses for both the designer and user tasks. Each point corresponds to a statement judgment in a given trial (normalized to $[0, 1]$), and the dashed line indicates perfect agreement ($y = x$). Each dot represents an average probability measure for a given trial. For designer game, each dot represents the probability of assigning a key in a tray. If the trial has N keys and M trays, that would correspond to a total of N x M measures. For user game, each dot represents a statement rating from each trial. If the trial has N keys and M doors, that would correspond to a total of N x M ratings.

For the **designer phase**, the literal designer baseline shows only a modest correspondence with human placements ($r = 0.316$, 95% CI $[0.139, 0.474]$). In contrast, the pragmatic designer model yields a substantially stronger fit ($r = 0.954$, 95% CI $[0.934, 0.968]$), capturing designers' tendency to make placements that are informative for the user rather than solely efficient.

For the **user phase**, both models predict human judgments

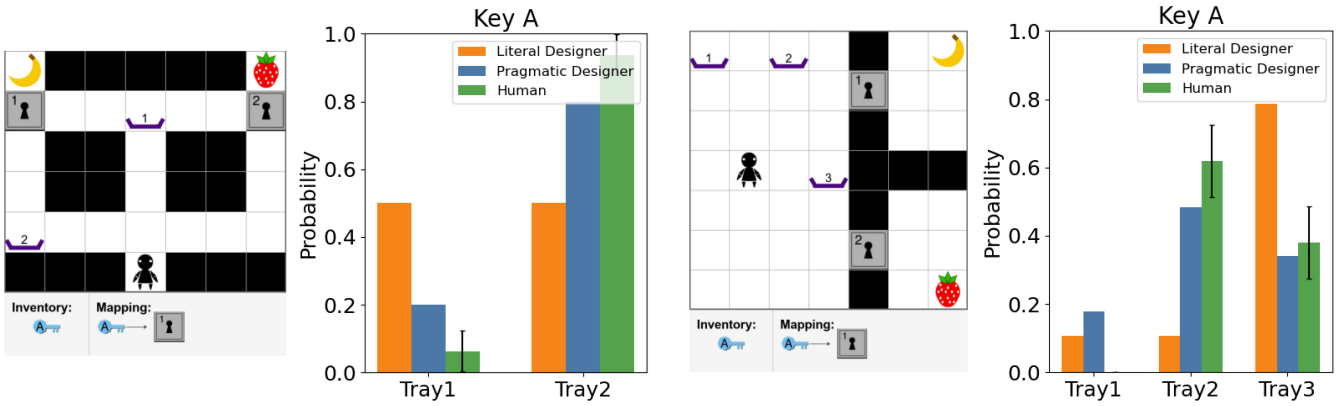


Figure 5: Qualitative designer examples comparing model predictions and human judgments. In each subplot, the map on the left shows the scenario, with the key available to the designer in the inventory and the ground truth mapping. Then the bar plot on the right shows the model prediction and averaged human response. We averaged all human participants' decisions to compute the probability of putting the key in each tray.

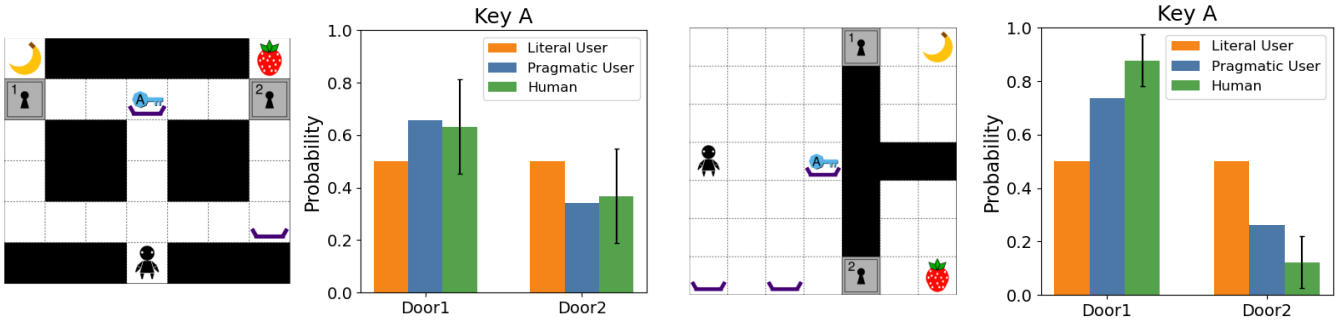


Figure 6: Qualitative user examples comparing model predictions and human judgments. On the left of each subplot shows a map designed by a human participant in the designer game. On the right, each user sees the map and infer which door does the key unlock.

well. The literal user achieves a strong fit ($r = 0.906$, 95% CI $[0.857, 0.939]$), while the pragmatic user further improves performance ($r = 0.949$, 95% CI $[0.922, 0.967]$), better matching the graded confidence users express when interpreting ambiguous placements.

Qualitatively, the literal user plot exhibits a pronounced vertical streak (many predictions concentrated at the rating of 0.5), even though human responses span a wide range. This pattern is consistent with the literal model producing discretized or tied posteriors when multiple doors are at similar distances (or when the softmax yields near-equal utilities), leading to the same predicted probabilities for many possible key-door mappings.

By comparison, the pragmatic user plot places substantially more mass near the diagonal, while also preserving the empirical tendency toward extreme judgments, indicating that both people and the pragmatic model often reach confident conclusions when placements are highly diagnostic.

To further illustrate these patterns, Figures 5 and 6 show qualitative examples of human judgments alongside model predictions.

Designer Phase Case Studies

Figure 5 illustrates two designer-trials. The left example shows a simple case where there are two purple trays. Tray 1 is equal distance between door 1 and door 2, and both trays are equal distance from door 1. The designer has Key A which unlocks door 1. We find that all human participants assigned the key to tray 1, which is predicted by the pragmatic designer model. This is because Tray 2 is more informative than Tray 1 to a user and thus have more communicative utility. In contrast, a literal designer is indifferent between Tray 1 and Tray 2 since they are equally efficient.

On the right, we show an example where designer needs to balance effiency and informativeness. In this example, the designer needs to allocate a key that unlocks door 1. Tray 3 is the most efficient option for door 1. A literal designer model therefore assign high probability of placing the key in Tray 3. However, Tray 3 is equal distance from both door 1 and 2, which makes it an option with low informativity. Tray 2, on the other hand, is slightly more inefficient but more informative. Both the pragmatic designer model and human participants show a lot more uncertainty between Tray 2 and

Tray 3, indicating a tension between balancing efficiency and informativity in pragmatic design.

User Phase Case Studies

Figure 6 illustrates two representative user-trials where the same spatial arrangement can support different inferences depending on whether the user assumes the placement is merely efficient (literal) or intentionally informative (pragmatic).

In the left example, Key A is placed in a tray adjacent to a corridor that separates the two doors. Under the literal user model, which relies primarily on spatial proximity, key A is approximately equidistant from door 1 and door 2, producing near-indifference (roughly 0.5/0.5; orange bars). Human responses, however, strongly favor door 1 (green bar near 0.9), suggesting that users interpret the placement as an intentional cue rather than as a purely geometric signal. The pragmatic user partially captures this effect (blue bars), assigning higher probability to "key A unlocks door 1" by reasoning that a cooperative designer would place a key so as to disambiguate the intended door. More generally, pragmatic inference can break symmetries in layouts where distance-only heuristics yield tied posteriors.

In the right example, the key is placed in the center tray, which is equal distance from both doors. Similar to the left example, the literal user interprets it to be equally likely to unlock door 1 and door 2. In contrast, both humans and the pragmatic user model assign higher probability to door 2, because if the key is supposed to unlock door 1, a pragmatic designer would have put the key in the left trays.

Failure case

While the pragmatic designer model achieved high correlation against human placement decisions, we also noticed a few outliers in the scatterplot (Figure 4).

The outliers are mainly due to one scenario shown in Figure 7. In this example, some human participants acted more like a literal designer, finding tray 1 to be an optimal place for the key that unlocks door 2. On the other hand, the pragmatic designer assigns equal utility for tray 2 and tray 3. This is because they are equally optimal and informative. We suspect that human participants may find tray 2 to be less optimal because it's farther away from door 2 and potentially more costly to visit (people may prefer walking diagonal paths).

Discussion and Conclusion

Our results support the view that designers do more than optimizing for efficiency: they place cues in ways that anticipate a user's downstream inferences. The pragmatic designer captures this strategic behavior substantially better than the literal baseline, suggesting that informativeness about hidden structure is an important component of design utility.

On the user side, the pragmatic user provides a consistent improvement over the literal distance-based model, even though the literal model already fits well. The qualitative structure of the scatterplots highlights a key difference:

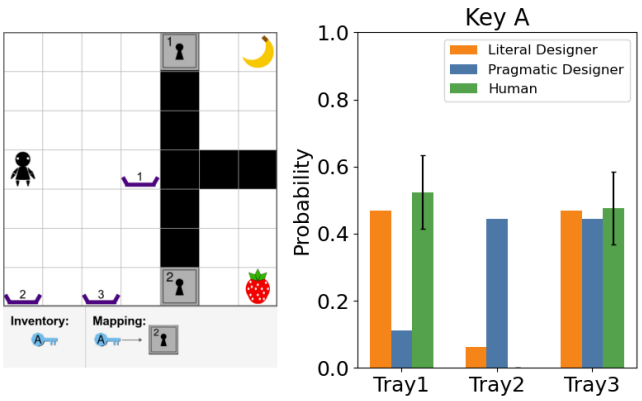


Figure 7: An example scenario where the pragmatic designer model does not match human judgment.

distance-based inference can collapse to tied or discretized predictions in symmetric layouts, whereas pragmatic inference smooths these ties by reasoning about why a cooperative designer would select particular placements. This aligns with the intuition that people interpret novel designs not only through perceptual cues, but also through assumptions about intentional, helpful design.

Our study is not without limitations. Our current formulation abstracts away several factors that likely matter in real artifacts, including heterogeneous user priors, learning over repeated interactions, and richer cost functions (e.g., physical effort, search, and error recovery). In addition, our evaluation uses a simplified grid-world task and a limited set of layouts, which may not capture the full complexity of everyday artifact understanding.

Future work may expand the space of communicative design actions beyond key placement (e.g., grouping, labeling, visual salience, or constraints that change the action set) and to study how multiple cues are integrated. We may consider more naturalistic design constraints such as physical laws. For instance, certain parts of the object may have physical affordances, which can affect the overall design (e.g., plane design needs to consider weight balancing for aerodynamics).

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Lawrence Erlbaum Associates.

Chandra, K., Li, T., Tenenbaum, J. B., & Ragan-Kelley, J. (2023). Inverse inverse planning [Manuscript].

Chen, A. M., Palacci, A., Vélez, N., Hawkins, R. D., & Gershman, S. J. (2024). A hierarchical Bayesian model of adaptive teaching. *Cognitive Science*, *48*(7), e13477.

Dragan, A. D., & Srinivasa, S. S. (2013). Generating legible motion. *Robotics: Science and Systems*.

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S. S., Griffiths, T. L., & Dragan, A. D. (2019). Pragmatic-pedagogic value alignment. *Robotics research: the 18th international symposium Isrr*, 49–57.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Gibson, J. J. (2014). The theory of affordances:(1979). In *The people, place, and space reader* (pp. 56–60). Routledge.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Hadfield-Menell, D., Dragan, A. D., Abbeel, P., & Russell, S. (2017). The inverse reward design problem. *Advances in Neural Information Processing Systems*.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. D. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*.

Ho, M. K., Littman, M. L., Cushman, F., & Austerweil, J. L. (2018). Effectively learning from pedagogical demonstrations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *40*, 505–510.

Lopez-Brau, M., & Jara-Ettinger, J. (2023). People can use the placement of objects to infer communicative goals. *Cognition*, *239*, 105524. https://doi.org/10.1016/j.cognition.2023.105524

Nielsen, J. (1993). *Usability engineering*. Morgan Kaufmann.

Norman, D. A. (2013). *The design of everyday things* (Revised and expanded edition). Basic Books.

Royka, A., Chen, A., Aboody, R., Huanca, T., & Jara-Ettinger, J. (2022). People infer communicative action through an expectation for efficient communication. *Nature Communications*, *13*, 4160. https://doi.org/10.1038/s41467-022-31716-3

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 105–140). John Wiley & Sons.

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems*, *33*, 19238–19250.

Zhi-Xuan, T., Ying, L., Mansinghka, V., & Tenenbaum, J. B. (2024). Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2094–2103.