From chopped to cooked: Design and inference in physical environments

Justin Yang, Lionel Wong, Judith E. Fan, Tobias Gerstenberg
{justin.yang, liowong, jefan, gerstenberg}@stanford.edu
Department of Psychology, Stanford University

Abstract

Physical spaces are often designed to support specific uses. But how do people create such environments, and how do users infer their intended function? We propose that design and inference about design are complementary processes, grounded in a capacity to mentally simulate goal-directed actions. We tested this using "Overcooked"-style kitchens where participants either judged what a kitchen was designed for (Study 1) or designed kitchens for cooks with varying goals and beliefs (Study 2). In Study 1, participants inferred that kitchens were designed for tasks the layout made easier to complete, consistent with the prediction of a simulation-based computational model. In Study 2, participants made designs that helped cooks efficiently complete their task, adjusting their choices when cooks faced uncertainty about which task to perform. Together, these findings point towards a study of design as a cognitive activity grounded in the same mechanisms that support planning and social reasoning.

Keywords: design; mental simulation; social cognition; Bayesian inference; inverse planning

Introduction

Imagine opening a new restaurant. How would you design the kitchen? If the restaurant If you need to prepare a buffet, you might arrange the equipment to support a wide variety of dishes. But if you were opening an Italian restaurant, you might specialize certain areas for specific tasks—for example, positioning flour and water within easy reach to streamline pasta-making. The same logic applies beyond restaurants. Suppose you are an urban planner designing an airport. Anticipating that travelers may run late, you might design wide corridors with clear signage so that people do not run into each other or get lost.

Designing Environments

The study of environment design has a long history in architecture and urban planning, where it has given rise to many intuitions about how to make an environment "well designed". For example, the notion that *form follows function*—that good design responds systematically to human needs—has received considerable attention in design theory and practice (Corbusier, 1923; D. Norman, 2013). In cognitive science, a smaller body of work has examined how people rearrange objects to reduce visual search times and facilitate foraging behaviors (Gray et al., 2006; Solman & Kingstone, 2017). While these studies demonstrate that people are sensitive to the relationship between spatial structure and task efficiency, so far no unifying computational framework has been developed that makes quantitative predictions about how people structure environments to support their goals.

Inferring Intent from Design

If environments reflect intentional design, people who recognize this should be able to infer their intended use and act accordingly. Consider a chef who has just been hired at a new restaurant. Walking into a kitchen where pasta-making equipment is prominently positioned and easily accessible, they will likely infer the restaurant's specialty. Similarly, a traveler navigating an unfamiliar airport can see that the wide corridors and prominent signage were placed deliberately, and trust these features to guide their path rather than wandering at random. Among design practitioners, this communicative dimension is called *legibility*: the idea that environments should make their function apparent to users (Montgomery, 1998; D. A. Norman, 1986).

In cognitive science, recent work has examined how people use Theory of Mind to extract social information from physical traces in the environment—for instance, inferring from objects in front of a doorway that others should not enter (Jara-Ettinger & Schachner, 2024; Lopez-Brau & Jara-Ettinger, 2023; Teo et al., 2025). Yet little work has examined how people reason about durably designed environments like kitchens and airports, which exist not just to *communicate* information about their use, but also to *support* users' goals.

The Present Research

In this work, we consider design and inference about design as complementary processes, grounded in broader capacities for reasoning about agents' beliefs, goals, and plans. We hypothesize that people design environments by reasoning about the goals and epistemic states of their users. We explore these ideas using kitchen environments inspired by the video game *Overcooked* (Ghost Town Games, 2016), in which cooks navigate gridworld kitchens to prepare dishes (Figure 1).

In Study 1, we ask whether people can infer what a kitchen was designed for from its spatial arrangement. In Study 2, we investigate how people design kitchen layouts for particular purposes, taking into account the goals and beliefs of the user. We find that people's inferences and design decisions are largely captured by a model grounded in planning, with some evidence that designers also consider users' beliefs.

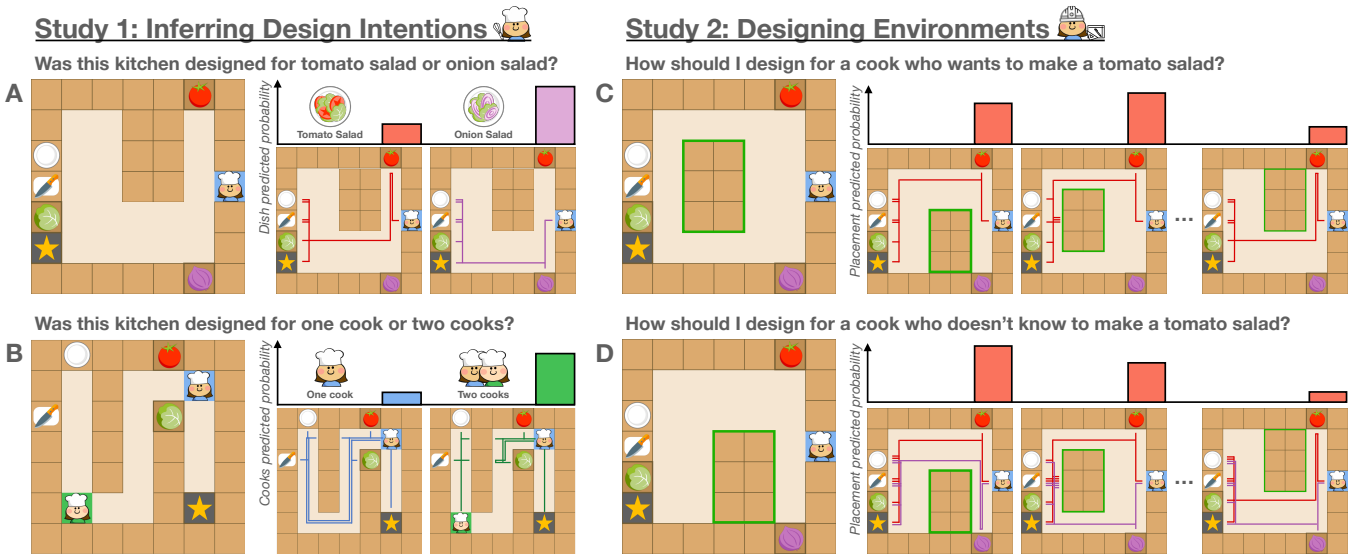

Figure 1: **Overview of Approach**. Each panel shows a kitchen with paths illustrating a cook's planning; bar plots show illustrative model predictions. (A)–(B) *Inference task*. Participants inferred whether kitchens were designed for a particular dish or number of cooks; the model assigns higher probability to whichever option requires fewer steps. (C)–(D) *Design task*. Participants placed furniture to help cooks who knew or did not know which dish to make; possible furniture placements are shown. In (C), the model assigns higher probability to placements that help the cook complete the dish in fewer steps; in (D), the model additionally considers whether placements signal which dish the cook should make.

Study Environment

In the *Overcooked* environment, cooks attempt to complete one of two possible recipes in as few steps as possible. Each kitchen is a 7×7 grid containing several functional objects: *ingredient dispensers* that supply tomatoes, onions, and lettuce; *cutting boards* where ingredients can be chopped; a *plate*; and a *delivery window* (marked with a star) where completed dishes are submitted. Cooks navigate the grid by moving in the four cardinal directions. They cannot walk through grid cells with counters or another cook. They cannot put items down on floors, and can carry only one item at a time.

Cooks complete a recipe by gathering required ingredients, chopping each one, combining them on a plate, and delivering the plated dish. The two recipes differ only in one ingredient: *tomato salad* (chopped tomato + chopped lettuce) and *onion salad* (chopped onion + chopped lettuce). Submitting the wrong dish resets the kitchen, forcing the cook to start over.

Computational Model

We develop a computational framework for modeling how people infer design intent and design environments for others.

Simulation model

We formalize the cook's planning as a Markov Decision Process (MDP; Bellman, 1957). A kitchen configuration $\mathcal{M}$ defines an MDP in which the cook takes actions to complete a dish in as few steps as possible. Let $C_{\mathcal{M}}(g)$ denote the minimum number of steps required to complete goal g in configuration $\mathcal{M}$. We estimate this using a hierarchical planner that decomposes recipes into subtasks and greedily selects the next subtask based on which has the lowest cost (see Wu et al., 2021, for full formalism).

When the kitchen has two cooks, we consider two models of multi-agent coordination. In the *independent* planner, each cook plans greedily, treating the other as a moving obstacle. In the *coordinating* planner, cooks use inverse planning to infer each other's current subtask and avoid duplicating effort (see the 'Bayesian Delegation model' in Wu et al., 2021).

Modeling inference about design intent

An observer who encounters a designed environment can infer what it was designed for by considering which tasks it makes easier to complete. If a kitchen makes tomato salad easier to prepare than onion salad, an observer should infer it was designed for tomato salad. We model this inference as a softmax over negative step counts:

$$P_{\mathcal{M}}(g) \propto \exp\big(-\lambda \cdot C_{\mathcal{M}}(g)\big), \tag{1}$$

where g indexes possible goals (e.g., tomato vs. onion salad) and λ controls how strongly the inference tracks efficiency differences.

Modeling environment design

In the design task (Study 2), participants place a 2×3 furniture block at position k within a base layout ℓ, yielding 12 possible configurations; we denote the resulting configuration as $\mathcal{M}_{\ell,k}$. We model designers as selecting placements according to a softmax over utilities:

$$P(k \mid \ell, g) \propto \exp\big(\beta \cdot U_{\mathcal{M}_{\ell,k}}(g)\big), \tag{2}$$

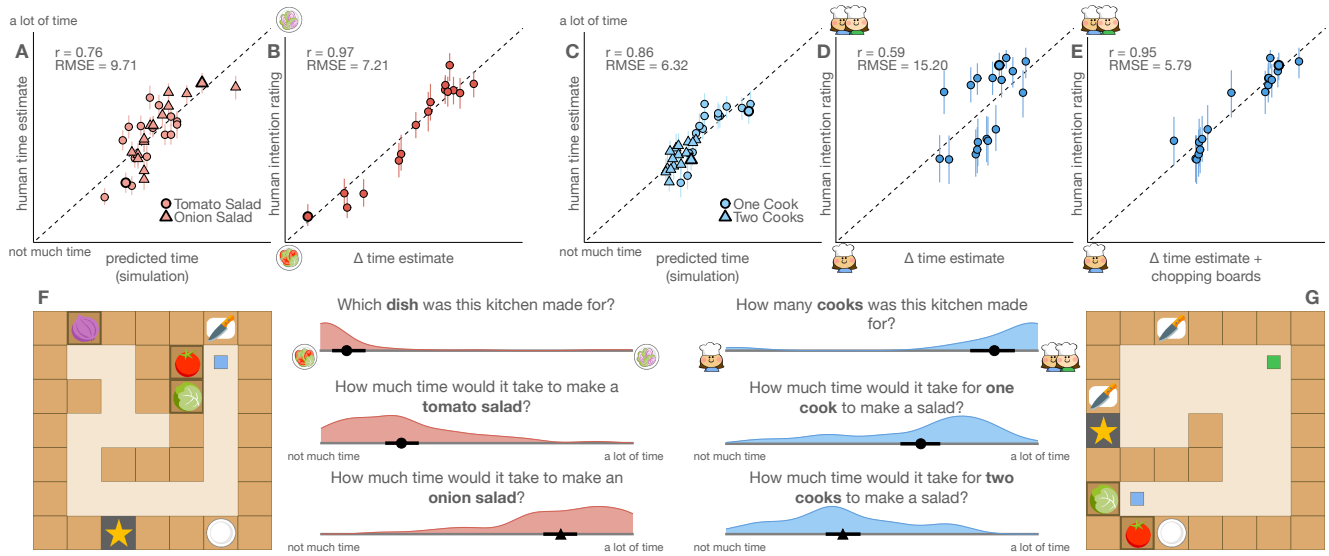

Figure 2: **Inference Study Results**. Each scatterplot shows one point per kitchen layout; y-axes show participants' mean ratings and x-axes show the corresponding model predictions. Error bars show 95% CIs. Bolded points correspond to the example trials in (F)–(G). (A)–(B) 'Dish condition'. Time estimates correlated with the planner's predicted step counts (A), and intention inferences tracked relative efficiency (B). (C)–(E) 'Cooks condition'. Time estimates correlated with model predictions (C), but intention inferences were only moderately predicted by relative efficiency alone (D); adding the number of cutting boards improved fit (E). (F)–(G) Example trials from the dish (F) and cooks (G) conditions, with response distributions shown.

where g is the intended goal and β controls how strongly participants prefer high-utility placements. When the cook knows which dish to make, the designer only needs to select placements that minimize step costs, giving the *Oracle* model: $U_{\mathcal{M}_{\ell,k}}(g) = -C_{\mathcal{M}_{\ell,k}}(g)$.

Trading off efficiency and communicativity. Consider a cook who does not know which dish to make and attempts the wrong one first. That cook incurs the full cost of completing the wrong dish, then must start over and complete the correct one. A designer who anticipates this possibility should evaluate each placement not only by how efficiently the intended dish can be completed, but also by how likely the cook is to attempt the wrong dish first. We formalize this with a utility function that captures both considerations:

$$U_{\mathcal{M}_{\ell,k}}(g) := -C_{\mathcal{M}_{\ell,k}}(g) - \left(1 - P_{\mathcal{M}_{\ell,k}}(g)\right) \cdot C_{\mathcal{M}_{\ell,k}}(\neg g), \quad (3)$$

where $\neg g$ denotes the alternative dish and $P_{\mathcal{M}_{\ell,k}}(g)$ is the probability the cook attempts the correct dish first (Equation 1). The first term captures the *efficiency* of completing the intended dish. The second term captures the *communicative cost*: the expected wasted steps if the cook misinfers and must restart. This is our *Full* model.

Alternative models. We compare the Full and Oracle models against two ablations. The *Inference-Ablation* model assumes the cook guesses randomly, so $P_{\mathcal{M}_{\ell,k}}(g) = 0.5$. The *Efficiency-Ablation* model retains the cook's efficiency-based inference but removes sensitivity to step costs: $U_{\mathcal{M}_{\ell,k}}(g) = P_{\mathcal{M}_{\ell,k}}(g)$.

Study 1: Inferring Design Intent

If environments are designed to support specific uses, observers should be able to infer what a space is for by noticing which tasks it makes easier. We tested this hypothesis by collecting two kinds of judgments. One group estimated how long different tasks would take in each kitchen; another judged what each kitchen was designed for. We predicted that intention inferences reflect assessments of relative task efficiency, as formalized in our computational model (Equation 1). We examined this in two contexts: inferring which dish a kitchen was designed for (tomato vs. onion salad) and how many cooks it was designed for (one vs. two)[1].

Methods

Participants Participants were recruited via Prolific. They self-reported gender by selecting from the options *male*, *female*, *non-binary*, and *other*. A total of 200 participants (*age*: median = 39, range = 20–88; *gender*: 102 female, 96 male, 2 non-binary; *race*: 17 Black/African American, 14 Asian, 1 American Indian/Alaska Native, 7 multiracial, 156 white, 5 other) completed the study. Participants were evenly assigned into one of four between-subjects conditions in a 2 (dish vs. cooks) × 2 (time-estimation vs. inference) design. All participants were native English speakers residing in the US. Participants were paid $2.40 for an estimated 12 minutes

[1]Materials, preregistrations, and data are available here. Experiments were coded using `jsPsych` (De Leeuw et al., 2023) and stored via `DataPipe` (de Leeuw, 2024). AI tools were used in the preparation of the materials.

to complete the study in the time-estimation conditions, and $3.00 for an estimated 15 minutes to complete the study in the inference conditions (*mean completion time*: 13.0 mins).

Stimuli We created 18 kitchen layouts for each condition. In the 'dish condition', kitchens contained one tomato, one onion, and one lettuce dispenser, allowing preparation of either salad. In the 'cooks condition', kitchens contained only a tomato and lettuce dispenser, and showed starting positions for two possible cooks. Layouts were hand-crafted to span a range of design intentions, and varied in the number of interior countertops (4, 6, or 9) and cutting boards (1 or 2).

Procedure Participants read instructions describing the Overcooked environment and task. Participants in the time-estimation conditions were asked to "estimate how long it would take to make a dish in each kitchen". Participants in the inference conditions were asked to "judge how each kitchen was designed to be used". In the 'dish condition', participants were told that some kitchens were designed for a cook to make an *Onion Salad* while others were designed for a *Tomato Salad*. In the 'cooks condition', they were told that some kitchens were designed for a cook to work alone while others were designed for two cooks to collaborate. Participants then completed a comprehension check containing questions about the environment mechanics. Incorrect responses returned them to the instructions. All participants passed the comprehension check, though some required multiple attempts (*mean failed attempts*: 1.0, *range*: 0–7).

Participants viewed 18 kitchen layouts in randomized order. On each trial, they saw a static image of the kitchen with the cook's starting position marked. In the 'cooks condition', a blue square marked where one cook would start, and a green square marked where the second cook would start; participants thus reasoned about the same kitchen with either one cook (starting at the blue square) or two cooks (starting at the blue and green square). Participants responded using sliders to provide either time estimates or intention ratings (Figure 2F–G shows example trials with response formats).

Results

For each kitchen layout, we compared the mean time estimates and the mean intention ratings across participants. We asked whether participants' time estimates aligned with our simulation model, and whether their judgments about what each kitchen was designed for tracked the relative efficiency of completing each task.

Inferring what dish a kitchen was designed for In the 'dish condition', participants judged whether each kitchen was designed for making a tomato salad or an onion salad (Figure 2F). We first compared how well our hierarchical planner, which approximates the minimum number of steps an agent would need to complete each dish, captured participants' time estimates. Simulated step counts were correlated with participants' average time estimates for both tomato salad ($r = 0.68$, $p = .004$) and onion salad ($r = 0.82$, $p < .001$), suggesting that participants' time estimates aligned with those of a rational planner (Figure 2A).

We then asked whether these time estimates predicted participants' intention inferences. For each kitchen, we computed two quantities: the average rating of which dish it was designed for, and the average difference in time estimates between dishes. We then fit a linear regression predicting intention inferences from the difference in time estimates. Intention inferences were strongly predicted by the difference in time estimates ($r = 0.97$, $p < .001$; Figure 2B): kitchens judged to be relatively *faster* for making a specific dish were also judged as *designed* for making that dish. When predicting participants' intention inferences using the *planner's* relative step counts, we find a similar pattern ($r = 0.75$, $p < .001$).

Inferring how many cooks a kitchen was designed for In the 'cooks condition', participants judged whether each kitchen was designed for one cook working alone or for two cooks collaborating to make a tomato salad (Figure 2G). For one cook, participants' time estimates were correlated with the number of steps the model required ($r = 0.78$, $p < .001$; Figure 2C). For two cooks, participants' time estimates were better captured by the coordinating planner ($r = 0.70$, $p = .003$) than the independent planner ($r = 0.65$, $p = .006$), so we used the former for subsequent analyses.

Participants' judgments of how many cooks a kitchen was designed for were correlated with their time estimates—specifically, the perceived advantage of two cooks over one—but this correspondence was weaker than in the dish condition ($r = 0.59$, $p = .010$; Figure 2D). However, we also explored whether participants drew on a simpler cue: the number of cutting boards in each kitchen, as a kitchen with two cutting boards might already suggest a design for two cooks, without needing to granularly simulate cooks acting in the kitchen. Consistent with this possibility, intention ratings were well-predicted by a regression model combining both the difference in time estimates and the number of cutting boards as linear predictors ($r = 0.97$, $p < .001$; Figure 2E). This exploration suggests that participants combined their assessments of relative task efficiency with this heuristic cue when inferring how many cooks each kitchen was designed for.

The planner's relative step counts were similarly not a strong predictor of participants' intention inferences on their own ($r = 0.44$, $p = .067$). However, combining model predictions with the number of cutting boards yielded a strong fit ($r = 0.95$, $p = .002$), mirroring the pattern observed with participants' own time estimates.

Study 2: Designing Environments

Study 1 showed that observers can infer what a kitchen was designed for by assessing which tasks it makes easier. We now ask how people *design* environments for others to use. Unlike observers, who need only assess what an environment supports, designers must anticipate how users will act—a task that requires representing not just physical constraints but also users' mental states. When users know their goal, de-

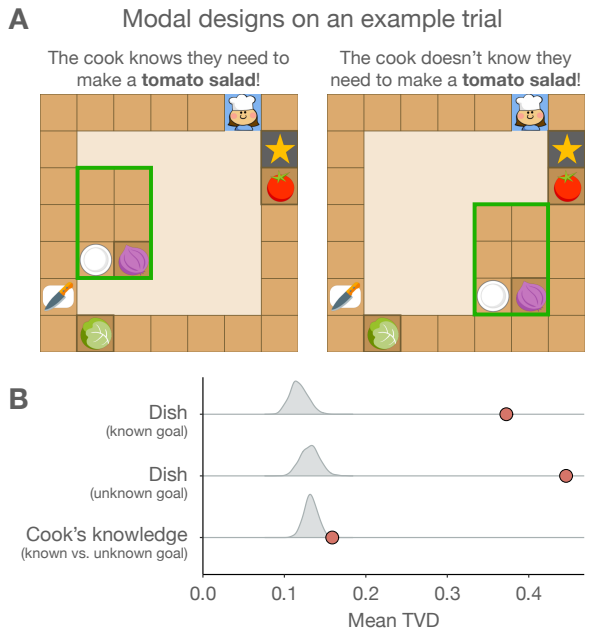

Figure 3: **Design Study Results**. (A) *Modal placements on an example trial*. In the known-goal condition, participants placed the furniture on the left-hand side. In the unknown-goal condition, participants blocked the onion, ensuring the cook could only make the intended dish. (B) Permutation tests comparing placement distributions (gray = null distribution; red = total variation distance (TVD) between placement distributions). Participants placed furniture differently depending on the target dish (top two rows) and, albeit much less so, depending on whether the cook knew which dish to make (bottom row).

sign can focus on efficiency; when users are uncertain, the environment must also help them figure out *what* to do.

To test whether designers are sensitive to users' epistemic states, we manipulated what participants believed about the cook who would use their kitchen. In the 'known-goal' condition, the cook would know which dish to make. In the 'unknown-goal' condition, the cook would not know which dish to make and would have to infer it from the environment. If designers consider users' uncertainty, we should observe different patterns of design choices across conditions. Designers should select layouts that not only support efficient task completion but also signal which task to perform.

Methods

Participants Participants were recruited via Prolific. A total of 200 participants (*age*: median = 38, range = 18–69; *gender*: 112 female, 85 male, 2 non-binary, 1 other; *race*: 29 Black/African American, 23 Asian, 3 American Indian/Alaska Native, 1 Native Hawaiian/Pacific Islander, 5 multiracial, 139 white) completed the study, divided evenly across two between-subjects conditions: 'known-goal' and 'unknown-goal'. All participants were native English speakers residing in the US and were paid $3 for an estimated 15 minutes (*mean completion time*: 15.5 mins). Both studies reported here followed the Stanford University IRB protocol. All participants provided informed consent.

Stimuli We created 12 base kitchen layouts. Each layout had a 5×5 interior with fixed objects (delivery window, plate, ingredient dispensers, cutting board): some positioned around the perimeter, others attached to a movable furniture block. Participants chose where to place a 2×3 table within the interior; this table could occupy any of 12 valid positions. Each participant completed 24 design trials: all 12 layouts crossed with both target dishes (tomato salad and onion salad), presented in fully randomized order.

Procedure Participants first learned about the Overcooked environment through illustrated instructions. Critically, they were told that submitting the wrong dish causes the kitchen to reset, requiring the cook to start over. In the 'known-goal' condition, participants were told the cook would know which dish to make. In the 'unknown-goal' condition, participants were told the cook would not know which dish to make and would have to infer which one to try first. Participants then completed a comprehension check testing understanding of both the environment mechanics and the condition-specific manipulation (*mean failed attempts*: 0.4, *range*: 0–8).

Participants then completed three familiarization trials where they controlled a cook to prepare dishes. Trials ended upon successful delivery of the intended dish. In the 'known-goal' condition, participants saw the target recipe; in the 'unknown-goal' condition, they saw both recipes and guessed which to submit first. To ensure participants in the 'unknown-goal' condition understood the cost of submitting the wrong dish, we designed the final familiarization trial so that participants would *necessarily* submit the wrong dish first and experience restarting.

On each design trial, participants saw a base kitchen layout, both possible dishes (with the target dish highlighted), and the furniture block. They dragged the furniture around the interior and could reposition it freely before submitting. In the 'unknown-goal' condition, each trial included a reminder that the cook does not know which dish to make.

Model fitting and evaluation. We fit each computational model separately for each experimental condition. For the cook's inference (Equation 1), we fix $\lambda = 1$ to avoid double-fitting non-linear parameters in our models. We estimate a posterior distribution over the softmax sensitivity parameter β for each model using the `R` package `brms` (Bürkner, 2017). We then compute expected log predictive density (elpd) via approximate leave-one-out cross-validation, treating each placement decision as a separate observation.

Results

Participants made varied furniture placements: across 24 trials, the median participant chose 9 unique placements of the 12 possible (range: 2–12), and only 5.5% used three or fewer placements across the entire study.

Table 1: **Design Study Results**. Δelpd: difference in expected log predictive density from leave-one-out cross-validation, relative to best model (with standard error; more negative = worse). Top-1 Acc. (%): proportion of trials where people select the model's modal response. *Human*: split-half noise ceiling measuring human agreement. 95% CIs from bootstrap resampling over design problems.

Model	Δ elpd (*se*)	Acc. (%)
Known-goal		
Oracle	**0 (0)**	30.2 (23.6, 37.8)
Full	−98.4 (14.8)	23.9 (16.8, 30.9)
Inf. Abl.	−9.3 (5.3)	**32.2 (27.0, 38.1)**
Eff. Abl.	−1810.7 (34.2)	8.4 (5.1, 13.2)
Uniform	−1356.7 (19.4)	8.3 (8.3, 8.3)
Human	—	41.6 (39.8, 42.9)
Unknown-goal		
Oracle	**0 (0)**	25.2 (19.0, 32.6)
Full	−68.2 (14.9)	22.2 (16.5, 28.7)
Inf. Abl.	−40.3 (7.1)	**25.9 (20.5, 32.0)**
Eff. Abl.	−1540.5 (41.8)	7.7 (4.7, 11.2)
Uniform	−1097.9 (31.8)	8.3 (8.3, 8.3)
Human	—	40.1 (38.6, 41.0)

Designing for a cook who knows which dish to make In the 'known-goal' condition, participants' furniture placements differed reliably depending on which dish the kitchen was designed for. As an initial exploratory check, we computed the total variation distance (TVD) between placement distributions for tomato salad versus onion salad within each layout, testing significance via permutation of dish label within each kitchen (Figure 3B; TVD $= 0.373$, $p < .001$).

We then compared models using Δ elpd and top-1 accuracy (see Table 1). Cross-validated model comparison strongly favored the Oracle model over a Uniform baseline ($\Delta\text{elpd} = -1356.69 \pm 19.38$), confirming that participants valued designs that supported efficient task completion. However, top-1 accuracy did not clearly discriminate among models: Oracle, Full, and Inference-Ablation all had overlapping confidence intervals. This suggests that while models differed in how they distributed probability across placements, they often agreed on which placement was most likely. A split-half noise ceiling shows that humans were more consistent with each other than with any model, but the models capture a substantial portion of the systematic variance.

Designing for a cook who does not know which dish to make In the 'unknown-goal' condition, participants' furniture placements similarly differed depending on the target dish (TVD $= 0.446$, $p < .001$). Participants also made different design decisions across conditions, though this effect was smaller than the differences by dish (Figure 3B; TVD $= 0.159$, $p = .010$). This suggests that participants represented the cook's uncertainty over which dish to make, and that this influenced their design choices. Figure 3A illustrates this pattern: in the 'known-goal' condition, participants placed furniture to minimize steps, whereas in the 'unknown-goal' condition, participants blocked access to the onion dispenser, ensuring the cook could *only* make the intended dish.

However, model comparisons did not reveal the predicted pattern (Table 1). We hypothesized the Full model would outperform the Oracle when cooks must infer their goal, but cross-validated comparison favored the Oracle in both conditions ($\Delta\text{elpd} = -68.16 \pm 14.92$). As in the known-goal condition, top-1 accuracy showed overlapping confidence intervals among the top models, suggesting that the models converged on similar modal predictions despite differing in their full distributions. The Efficiency-Ablation model performed substantially worse ($\Delta\text{elpd} = -1472.35 \pm 51.92$), indicating that efficiency remained central to participants' design choices even when cooks faced uncertainty. Although participants reliably made different designs when cooks faced uncertainty, our models did not capture this difference.

Discussion

We asked whether design choices and inferences *about* design are grounded in reasoning about how agents act to achieve their goals. We found evidence for this in both directions. Designers created environments that helped users complete the intended task. When designers believed users were uncertain about their goal, they adjusted their designs accordingly. Observers inferred design intent by assessing which tasks an environment made relatively faster to complete. In both cases, a model that simulated efficient, goal-directed agents also captured patterns in people's judgments about *design*.

One limitation is that our design task did not clearly discriminate between models based on efficient action alone, and those that additionally reasoned about the communicative intent of a designer. This may have been due to an overly simple "design" setting, where efficient designs for completing a goal were often very similar or identical to designs that would have overtly "communicated" that goal. Follow-up work will explore more complex environments that can better distinguish between these different design intentions.

Our current work also focused on environments designed for users to efficiently complete their goals. But good design can serve many other functions. A classroom might sacrifice physical efficiency to prevent injuries or reduce distraction. A core tenet of real design practice is also that "good design is unobtrusive" (Lovell, 2011). This might mean that designers choose to invest mental effort so that users do not have to – so that knowing where to go, or what to do, seems transparent to the user (Gibson, 1979; Kirsh, 1996; Rubio-Fernandez et al., 2025). Future work can explore how designers build environments to shape how people *think* within them.

Finally, our paradigm treated design as a one-shot choice among discrete options. Real-world design is often iterative and involves ill-defined objectives (Goldschmidt, 1991; Rittel & Webber, 1973; Simon, 1969). A complete account of design will require studying how people explore and revise in underspecified, unbounded design spaces.

Acknowledgments

The authors would like to thank the anonymous reviewers for helpful feedback and comments. JY was supported by an NDSEG fellowship. LCW was supported by a Stanford Institute for Human-Centered Artificial Intelligence (HAI) fellowship. JEF is supported by NSF CAREER Award #2436199, NSF DRL #2400471, and awards from the Stanford Human-Centered AI Institute (HAI) and Stanford Accelerator for Learning. TG was supported by grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and from the Cooperative AI Foundation.

References

Bellman, R. (1957). A markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Corbusier, L. (1923). *Towards a new architecture*. Courier Corporation.

De Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

de Leeuw, J. R. (2024). DataPipe: Born-open data collection for online experiments. *Behavior Research Methods*, *56*(3), 2499–2506.

Ghost Town Games. (2016). Overcooked [Published by Team17].

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Goldschmidt, G. (1991). The dialectics of sketching. *Creativity research journal*, *4*(2), 123–143.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological review*, *113*(3), 461.

Jara-Ettinger, J., & Schachner, A. (2024). Traces of our past: The social representation of the physical world. *Current Directions in Psychological Science*, *33*(5), 334–340.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, *4*(3-4), 415–452.

Lopez-Brau, M., & Jara-Ettinger, J. (2023). People can use the placement of objects to infer communicative goals. *Cognition*, *239*, 105524.

Lovell, S. (2011). *As little design as possible: The work of Dieter Rams*. Phaidon.

Montgomery, J. (1998). Making a city: Urbanity, vitality and urban design. *Journal of urban design*, *3*(1), 93–116.

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

Norman, D. A. (1986). Cognitive engineering. In *User centered system design* (pp. 31–62). CRC Press.

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, *4*(2), 155–169.

Rubio-Fernandez, P., Berke, M. D., & Jara-Ettinger, J. (2025). Tracking minds in communication. *Trends in Cognitive Sciences*, *29*(3), 269–281.

Simon, H. A. (1969). *The sciences of the artificial* (1996th ed.). MIT Press.

Solman, G. J., & Kingstone, A. (2017). Arranging objects in space: Measuring task-relevant organizational behaviors during goal pursuit. *Cognitive Science*, *41*(4), 1042–1070.

Teo, V., Wu, S. A., Brockbank, E., & Gerstenberg, T. (2025). Leave a trace: Recursive reasoning about deceptive behavior. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.