Design Intention Inference for Virtual Co-Design Agents

Matthew V. Law, Amritansh Kwatra, Nikhil Dhawan, Matthew Einhorn, Amit Rajesh, Guy Hoffman
[mvl24,ak2244,nd353,me263,ar883,hoffman]@cornell.edu
HRC^2 Lab, Cornell University
Ithaca, New York, United States

ABSTRACT

We address the challenge of inferring the design intentions of a human by an intelligent virtual agent that collaborates with the human. First, we propose a dynamic Bayesian network model that relates design intentions, objectives, and solutions during a human's exploration of a problem space. We then train the model on design behaviors generated by a search agent and use the model parameters to infer the design intentions in a test set of real human behaviors. We find that our model is able to infer the exact intentions across three objectives associated with a sequence of design outcomes 31.3% of the time. Inference accuracy is 50.9% for the top two predictions and 67.2% for the top three predictions. For any singular intention over an objective, the model's mean F1-score is 0.719. This provides a reasonable foundation for an intelligent virtual agent to infer design intentions purely from design outcomes toward establishing joint intentions with a human designer. These results also shed light on the potential benefits and pitfalls in using simulated data to train a model for human design intentions.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Human-AI co-design, Intention recognition, Design agents

ACM Reference Format:
Matthew V. Law, Amritansh Kwatra, Nikhil Dhawan, Matthew Einhorn, Amit Rajesh, Guy Hoffman. 2020. Design Intention Inference for Virtual Co-Design Agents. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423861

1 INTRODUCTION

Intelligent virtual agents have the potential to aid human designers. If we think of the design process as a search through a space of potential solutions to a task [27], computational agents can search at a scale and with precision that outstrips any human designer. Still, humans possess the ability to reason abductively, which can allow them to navigate the ill-defined and highly contextual nature of most design tasks more efficiently than heuristic search methods. These unique but complementary capacities suggest the advantage of blending computational design with human intuition.

In this work, we propose a step toward realizing virtual co-design agents by examining how an agent might infer a human partner's design intentions with respect to a task. Specifically, we present the following contributions:

(1) A probabilistic model relating design intentions to outcomes.
(2) A search agent to simulate design data used to train the parameters of this model.
(3) An LSTM-FCN network [14] for predicting design intentions from objective traces.
(4) An interactive system to collect data from a multi-objective design task.
(5) Experimental results of intention inference using data logs from the interactive system.

We illustrate our approach using the civic design task of drawing voting districts in the United States. Every ten years, US states are required to re-draw geographic boundaries of voting districts. Drawing "fair" districts is a difficult design problem, as district boundaries can affect representation across interest groups, and there are usually trade-offs between well-intentioned measures of fairness. It is up to the district designer to balance these, and different intentions can lead to very different designs.

To study design intention inference in this context, we implemented a system for humans to design voting districts and visualize outcomes (Figure 1). We collected data of designers using this system and developed a computational system that tries to infer fairness-promoting design intentions from observed outcomes.

2 MOTIVATION

Our interest in peer collaboration in design tasks is inspired by recent work that attests the benefits of having humans and artificial intelligence (AI) play equal roles in co-creative tasks. We specifically focus on design intentions because joint intentions are critical to any collaborative effort [7], but may be difficult to establish for design tasks, which tend to be highly ill-defined [8, 24].

2.1 AI, Design, and Enactive Co-Creativity

2.1.1 What should an Intelligent Design Agent Do? Interactive virtual agents have not traditionally been the focus of applying AI to design. Schön delineates *functional equivalence*, where computational tools enhance specific elements of human design activity, from *phenomenological equivalence*, where agents emulate core human design activities [26]. Most research in AI-supported design hews to the perspective of functional equivalence. For example, systems help humans evaluate designs [25], answer queries about a design space [5], and make suggestions [31]. Interactive genetic

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423861

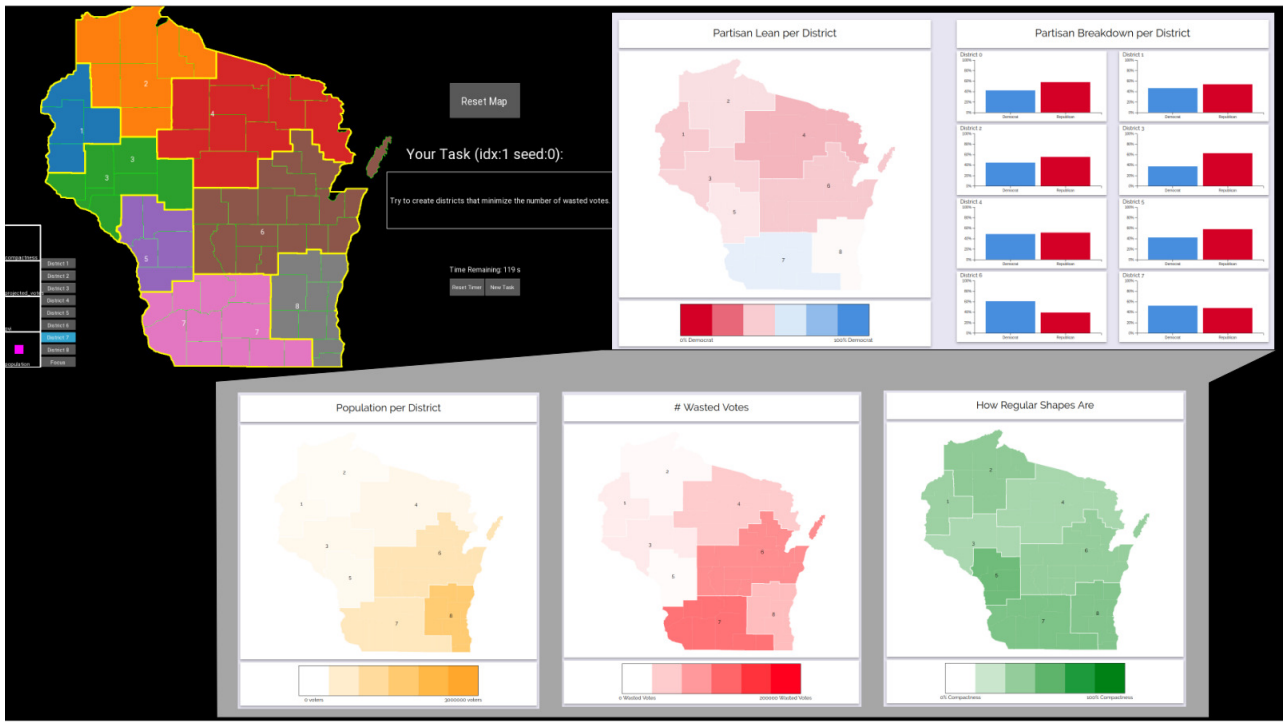

Figure 1: Users of our interface can partition a state into districts using an interactive map (top left). As they construct designs, they can visualize fairness-related outcomes as overlays on the state map accompanied by bar graphs (right and bottom).

algorithms generate new solutions on behalf of a human expert who guides the search by evaluating the designs it finds [4].

2.1.2 Enactive Co-Creativity with Computer Colleagues. Recently, Davis *et al.* have argued that both humans and agents can benefit from creative collaborations where agents play more human-like roles [10]. Enactive cognition posits that humans make sense of the world by interacting with it. Likewise, enactive agents can be designed to adapt and learn through improvisational, collegial interactions with human creators that enable more fluid co-creativity.

The future of AI design agents probably lies somewhere between support tools and virtual colleagues. Agents should address tasks that computers, and not humans, are good at. At the same time, design is highly social and benefits from the interaction of different "design worlds". Agents that emulate enough aspects of human designers to serve as creative foils may offer the best of both worlds.

2.2 The Role of Intentions in Collaboration

Virtual design partners must be able to read human design intentions. Joint intentions are well-established as prerequisite to collaborative effort. Bratman differentiates intentions from goals or desires as "intimately related to endeavoring and action" [7], not only motivating, but *controlling* what we do. This makes collaborative action without joint intentions difficult or impossible. This reality applies to human-agent collaboration. Allen *et al.* assert that negotiation of shared objectives is necessary for collaborative agents [2]. Similarly, Vernon *et al.* argue that the ability to perceive low-level and high-level intentions is essential in the design of cognitive robots [29]. In our own work, we observed collaboration breakdowns when interactions between a human and an AI agent surfaced different design intentions [17]. As discussed in the next section, we believe that the problem of intention inference is particularly challenging in the context of design tasks.

2.3 Ill-Definition and Design Intentions

Design tasks are a subset of ill-defined problems. Rittel and Webber coined the term "wicked problems" to describe the complexity and ambiguity that confront designers [24] and shape how they think [8]. Early attempts to formalize design (e.g. [1, 27]) required a task to be structured before it could be solved. As Simon points out, choices made about structure heavily influence solutions [27].

This ill-definition amplifies the importance of intention inference to collaboration. Effective collaborators must understand how each of their teammates interpret a task and intend to solve it. While it may be difficult for a designer to formulate or express their intentions a priori, humans develop the ability to infer others' intentions from actions at a young age [29]. As a first step towards collaboration, we ask how a virtual agent might read a human's design intentions from their choices as they explore a solution space.

3 RELATED WORK

Our work draws on two bodies of literature: computational models of how designers think, and intelligent systems inferring human intentions from their behavior.

3.1 Computational Models of Design Thinking

We build on a rich history of computationally modeling human design behavior, with a particular focus on the role played by design intentions. A full accounting of computational models of designing is beyond the scope of this paper; a comprehensive map of such models can be found in [30]. Models of design behavior can typically be broken down into prescriptive (e.g. [22]) and descriptive (e.g. [11, 27]), as well as stage-based (e.g. [27]) and process-based (e.g. [11]). In this work, we adopt a descriptive, stage-based model of designing. We are heavily influenced by sequential models of design in which the designer alternates between constructing a solution and reformulating the task [18]. However, we assume a static formulation of the task (the designer's intentions) for simplicity.

3.2 Inverse Reinforcement and Preference Learning

One way to understand human intentions is through a reward function that motivates how they act. Assuming a human seeks to maximize reward, the reward function can be inferred from their behavior. One approach, inverse reinforcement learning [21] (IRL) solves for a reward function given a known human expert's policy or sampled trajectories of behavior generated by this policy. Optimizing over an appropriate reward function is a critical factor in agent effectiveness, and, under some assumptions, IRL can avert the burden of reward engineering in agent design.

However, IRL canonically optimizes these reward weights assuming that the policy or observed behavior is optimal. When designing, however, even if the designer knows the optimal policy, they may not always adopt it, choosing to experiment with new ideas instead. Figuring out how to relax the optimality assumption for human behavior is an open problem in IRL, with some arguing that this may be impossible without strong normative assumptions [20].

Another approach to ease reward engineering, preference learning, learns a reward function that encodes human preferences offline, using supervised methods with human-labeled pairwise preferences [9]. However, our goal is not to learn implicit features that underlie a general set of human preferences, but rather how to differentiate between individual preferences for different designers in the context of how they explore a complex, unstructured task.

4 A MODEL OF DESIGN INTENTIONS

We formalize the design process as a Markov process in which the designer moves from one solution to another according to their intentions and what they have learned along the way.

4.1 Intentional Design as a Markov Process

A designer has beliefs about how design features cause outcomes and has intentions about the quality of these outcomes (Figure 2). To explore the design space, they modify a current solution or synthesize a new one, observing the changing outcomes as the design evolves. Each of these observations can influence the designer's beliefs and drive the next design change. This process continues until the designer is satisfied with the current solution. This formulation does not necessarily require that the designer operates rationally, either in how they perceive outcomes, update their beliefs, or synthesize a new design.

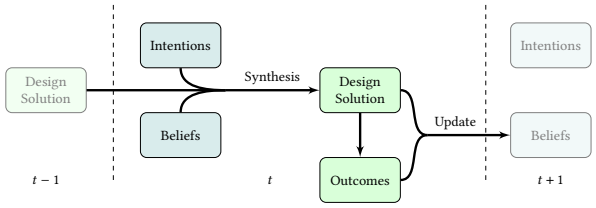

Figure 2: Each new design solution is synthesized from the current design based on the designer's intentions and beliefs about the design space. Beliefs are updated based on how the current solution affects observed outcomes.

4.2 A Probabilistic Graphical Model of Design

More concretely, let X be the space of design features associated with a task, and Y the space of measurable design outcomes, with $f : X \Rightarrow Y$ the mapping between the two. The designer typically does not know f explicitly but holds beliefs $b \in B$, over the space of possible functions f. The designer also intends to realize certain design outcomes; suppose that there is some set of all possible intentions, $I_{all} = \{i_0, i_1, ...\}$, where each intention targets some dimension of Y, for example, improving the voter efficiency of a district design. We represent the subset of intentions held by the designer, $I \subseteq I_{all}$, as a binary vector, where each component indicates the presence (1) or absence (0) in I of an intention in I_{all}.

This process can be represented using a Dynamic Bayes Network (Figure 3). Starting with some initial design, x_0, informed by their intentions, I, the designer observes y_0, and updates their beliefs about the design space, b_0. Based on x_0, y_0, b_0, and I, the designer constructs x_1, observes outcomes y_1, updates their beliefs to b_1, and so on, terminating the process when they are satisfied after some n steps. We are making the simplifying assumption that a designer's intentions do not change for the duration of a design session.

From the perspective of an agent observing the human designer, only the designs $x_0...x_n$ and corresponding outcomes $y_0...y_n$ are observable. If the agent wants to maintain joint intentions with the human, it has to infer I from sequences of designs, outcomes, and beliefs. In this paper, we simplify the model to include only the intentions, I, and outcomes, $y_0, ...y_n$ (shaded region in Figure 3). We remove $x_0, ...x_n$, as, depending on the complexity of the design feature space, observing how the actual design changes may be less informative than how the outcomes associated with those designs change. Additionally, we remove the model of designers' beliefs as these are neither observable, nor necessarily known to the designer. We discuss the cost of these simplifications in Section 7.

5 INFERRING INTENTIONS FROM OUTCOMES

To infer the probability that the designer has some intention i given a sequence of outcomes they designed, we use a neural network that approximates $P(i \in I | y_0...y_n)$. We frame this as a multi-label time-series classification problem, where each sequence of design outcomes is associated with one or more design intentions. Our approach to solving this multi-label problem is to convert it to a multi-class classification problem over the power set of I_{all}, as described in [28]. If we do not consider the possibility of a designer

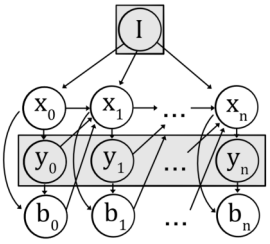

Figure 3: The design process described in Figure 2 can be represented as a Dynamic Bayes Net. Each design, x_t, is dependent on the previous design, x_{t-1}, and its outcomes, y_{t-1}, the designer's intentions, I, and the state of the designer's beliefs, b_{t-1}. The design and its outcomes influence the state of the designer's beliefs about the design space.

having no intentions, our problem thus becomes mapping a set of m design outcomes onto a probability distribution over $2^k - 1$ classes, spanning $k = |I_{all}|$ possible intentions.

5.1 Classifying Outcome Trajectories with LSTM-FCN

Drawing on Karim *et al.* [14], we use an LSTM-FCN network as a classifier. While fully convolutional neural networks can effectively learn structure in time series data, training one together with an LSTM has been shown to outperform the features learned by either approach independently [13]. Our architecture for the network is adapted directly from [14] and specified in Figure 4.

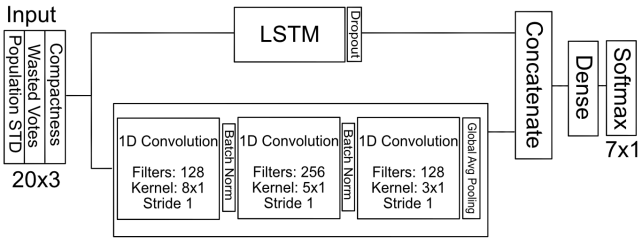

Figure 4: We classified design intentions from observed outcomes using an LSTM-FCN architecture adapted directly from [14]. The input to the network is a 20-step sequence of design outcomes; the output is a probability over the seven possible combinations of design intentions.

To augment our data and extract fixed-width observation sequences, we slid a window of twenty steps across the entire exploration trace. These sequences were passed through both a series of 1-D convolutions and an LSTM block. The outputs of the FCN and LSTM are passed to a final seven-dimensional softmax layer, where each node represents the probability that the sequence was generated by one of seven discrete binary intention vectors.

5.2 Data Collection

We trained and evaluated our model using trajectories of design outcomes observed while exploring the voting district design task described in Section 1. In this study, we chose to focus intentions on three district design outcomes: balancing the population between districts, improving voter efficiency by minimizing wasted votes, and maximizing the compactness of district shapes. A subset of design intentions in this context is a binary 3-vector of the form $I \in \{0, 1\}^3$, corresponding to these three intentions, respectively.

5.2.1 Collecting Human Design Exploration Data. We first asked humans to design for different sets of fairness design intentions and recorded the design outcomes of the solutions that they explored along the way. To this end, we built a custom redistricting interface for the US state of Wisconsin, described in the next section. We collected outcome traces from four members of our research team working on each of the seven combinations of intentions (*e.g.*, $[1, 0, 1]$, $[0, 1, 1]$) until they were satisfied. The intentions were presented in random order. In total, these traces contained design outcomes for 4826 design steps, an average of 689.4 steps per task (Table 1).

5.2.2 Distopia: An Interface for Voting District Design. Distopia is a system that allows either humans or virtual agents to explore different voting district designs, using a similar interface. Human designers interact with the system through a two-window graphical interface (Figure 1). On the control screen (top left in the figure), the human can read information about the design task, divide a map of the state into districts, and select different outcomes to visualize. The human is presented with a set of intentions in natural language. For example, *"Try to evenly balance the number of people each district has. Try to create districts that minimize the number of wasted votes. Try to create districts that are round and regularly shaped."* We also provide a timer and buttons to move on or reset the map and timer.

Designers draw districts by placing numbered markers on the map. Each marker allocates the space around it to a corresponding district ID. This is accomplished by performing a Voronoi decomposition of the map around the markers, discretized to the state's county boundaries. To realize complex district shapes, a designer can place multiple markers; partitions with the same ID are merged.

Outcomes are calculated and visualized for each design at the district level. The population of each district is summed over the counties that compose it. We use historical election data [15] to calculate the partisan lean for each district, summing over the votes for Republicans and Democrats in each county. A district's wasted votes are counted by taking the margin between the number of winning votes and half the votes in that district. Finally, we calculate the compactness of each district's shape as the ratio between its area and the area of a circle with the same perimeter as the district [23]. Users can overlay heatmaps for any of the outcomes on the current districts, or view the partisan breakdown of votes in each district.

We record each design outcome as a single aggregate score over all the districts in the state. These scores are the standard deviation of district populations, maximum number of wasted votes, and average compactness. The three design intentions map to minimizing each of the first two scores and maximizing the last, respectively.

5.2.3 Local Search as a Proxy for Human Design Exploration. The amount of data needed to train an intention classifier poses a stiff challenge in the context of designing. Design tasks are often one-off problems or rare events, such as the once-in-a-decade voter

DI	[0,0,1]	[0,1,0]	[1,0,0]	[0,1,1]	[1,0,1]	[1,1,0]	[1,1,1]
Steps	780	689	632	740	751	603	631

Table 1: Number of steps explored per set of design intentions (DI) in the human test data. The intentions are labeled by binary vectors, with indices mapping onto population balance, voter efficiency, compactness, respectively.

redistricting task. Moreover, they often take considerable time and effort, with parameters that do not easily generalize across contexts. As a result, models that require large amounts of data, such as neural networks, are difficult to train in the design context.

We approach this challenge by training on simulated data, using a search agent as a proxy for a human designer. After generating a model, we evaluate it on the human-generated test set. We discuss the promise and shortcomings of this approach in Section 7.

An ϵ-greedy local search agent serves as our proxy for a human designer. Exactly like the human designer, the agent partitions the state into districts by arranging district markers. The agent's design is initialized with a random collection of markers, constrained such that it has at least one and no more than five per district. For each agent run, the design intentions are static, and the agent evaluates the initial design using the aggregate outcomes associated with each intention in the predefined set. Outcomes are z-standardized to account for different scales, using the distribution of a 50,000-design random walk. Once it has calculated the quality of the current design, the agent samples potential modifications, each consisting of moving one district marker. It evaluates each change based on the metrics relevant to its design intentions, then chooses the best one, choosing something random instead with probability ϵ. We set ϵ at 0.8 with a decay of 10% per step, down to a minimum value of 0.1. To produce our training data, we ran this ϵ-greedy agent over the seven combinations of intentions for 130 episodes of 100 steps each. This resulted in a total of 73,710 sliding-window samples of 20-steps, as discussed below.

6 RESULTS

We trained the LSTM-FCN network on 70% of the simulated data, holding 30% out randomly for validation. Before training, we z-standardized the simulated data and augmented the data by sliding a window of 20 steps across each episode. We updated the network using cross-entropy loss and the Adam gradient descent method [16], with a learning rate of $1e{-}3$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The model fit the training data well, with a final training accuracy of 0.960 and validation accuracy of 0.945 after 50 epochs.

We then used the trained network to infer design intentions on the human-generated data set. We separately z-standardized and windowed the human data in the test set. Since the length of human exploration on each set of intentions varied, we randomly sampled from the windows to achieve a balanced set of 500 per class.

6.1 Metrics for Multi-Label Classification

To evaluate the performance of the multi-label classifier, we use two notions from Sorower [28]: *complete accuracy* and *partial accuracy*. Complete accuracy describes the rate at which the network correctly predicts the human's intentions with respect to all three outcomes (population balance, voter efficiency, and compactness), i.e. the binary representation (*e.g.,* $[1, 0, 1]$) of intentions. We also report how often the complete intention set was in the network's top two or top three most probable classes.

We are also interested in partial accuracy, i.e., the precision and recall for each individual intention. The precision with respect to any single intention (e.g., compactness) indicates how often the designer was actually trying to achieve compactness when the network predicted that they were, regardless of other intentions. Similarly, recall would tell us how often, if any particular intention was present, that it was predicted as one of the intentions.

6.2 Complete Accuracy

Our network achieved a complete accuracy of 0.313 over the seven subsets of intentions in the human exploration data. Since the test data was balanced over seven classes, this compares to a chance accuracy of 0.143. Taking into account class probabilities at the output layer, we further find that the network had a complete accuracy of 0.509 over its top two choices, and 0.672 over its top three. Figure 5 shows the confusion matrix for the class predictions and precision and recall for each class. Figure 6 plots the mean and per-class recall in the top-1, top-2, and top-3 probabilities, compared to the baseline chance recall for each type of test. These plots indicate that the classifier did much better at predicting certain subsets of intentions than others. The classifier tends to have higher recall and precision for classes that only contain one intention (blue in Figure 6), in particular population balance ($[1, 0, 0]$) and voter efficiency ($[0, 1, 0]$). While compactness alone ($[0, 0, 1]$) has lower recall, 92.5% of the false negatives for that class predict an intention to maximize compactness, suggesting higher partial accuracy. The classifier does particularly poorly on the multi-intention classes $[1, 1, 0]$ and $[1, 1, 1]$. That said, 72.4% of the false negatives for $[1, 1, 0]$ are split between the two corresponding single-intention classes ($[1, 0, 0]$ and $[0, 1, 0]$), and 72.0% of the false negatives for $[1, 1, 1]$ are split between the three classes with two intentions. These trends can be illustrated in the partial accuracy analysis.

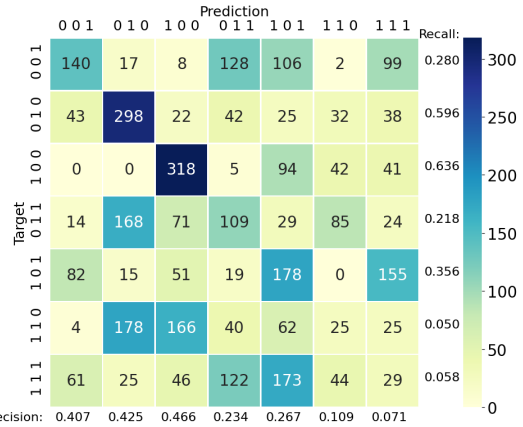

Figure 5: This confusion matrix shows prediction frequency for each subset of design intentions, represented as binary masks over [balance population, improve voter efficiency, maximize compactness]. For example, 101 represents intentions to balance population and maximize compactness.

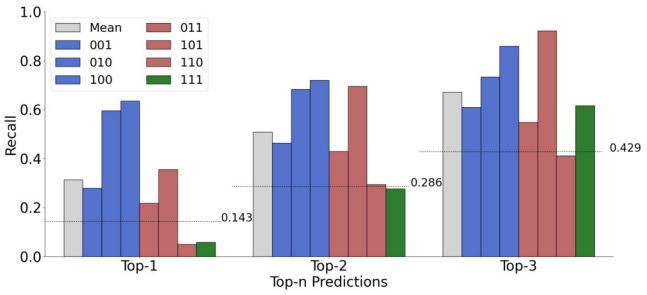

Figure 6: This plot shows the frequency with which the classifier predicted all three of the design intentions for each class, color-coded by the number of intentions, within the top one, two, and three most probable classes. The dotted lines show the threshold of random chance for each of these and the mean across classes for each group is shown in gray.

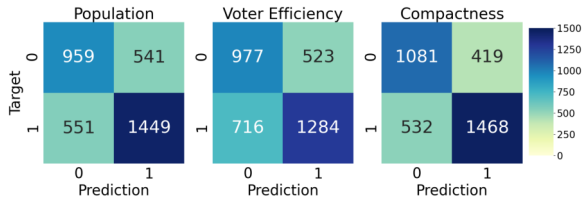

Figure 7: These confusion matrices show the frequency with which the model predicted the designer's intentions towards a specific outcome.

6.3 Partial Accuracy

On the level of individual intentions, we found that the network's predictions had a precision of 0.728 and recall of 0.725 (F1-score 0.726) for *balancing population*, precision 0.711 and recall 0.642 (F1-score 0.675) for *improving voter efficiency*, and precision 0.778 and recall 0.734 (F1-score 0.755) for *maximizing compactness*. Averaging across the labels produces overall precision 0.739 and recall 0.700, (F1-score 0.719). When the network predicted a human was designing for any one of the three intentions, it was correct, on average, 73.9% of the time, and it detected when the human was designing for an intention, on average, 70.0% of the time. Interestingly, in contrast to the class-level metrics, but in-line with our analysis above, the network actually predicts whether the human is maximizing compactness most reliably. Confusion matrices for each of the intention labels can be seen in Figure 7.

7 DISCUSSION

Results from our evaluation indicate that an agent observing a designer can infer their complete intentions over three objectives with better-than-chance (31.3% vs. 14.3%) accuracy, have them within its top two predictions with probability 50.9%, and detect individual intentions with circa 70% precision and recall. It does so by learning about the probabilistic intention-objective relationship using simulated data generated by an ϵ-greedy search agent.

These results can provide a basis for the development of intention-aware virtual design agents, but also highlight tradeoffs to consider, especially when using simulated design behavior as the basis for real-world intention inference. We discuss these topics below, as well as ways to improve our model of design intentions.

7.1 Toward Intention-Aware Co-Design Agents

Given the ability to infer a human's design intentions based on their design activity, how can a virtual agent use this information to be a useful co-designer? Perhaps the simplest way for an agent to operationalize its predictions is to adopt them itself. Our ϵ-greedy search agent, for example, could use predicted intentions to weigh outcomes when evaluating new designs. With that in mind, we note that our model, while better than random, did not reliably predict the human's intentions with respect to all three design outcomes at once. Adopting predictions that are correct 31.3% of the time may not yield design choices sufficiently aligned with a human partner to maintain collaborative effort. In contexts with more than three possible design intentions, this would be even more difficult.

A more nuanced approach of considering labels individually, however, seems promising. If a person intended to design for a specific outcome, our model identified it 70.0% of the time. Conversely, 73.9% of the time when it predicted an intention, the designer was actually designing for the associated outcome. A virtual co-design agent could marginalize across the softmax output of the neural network to extract a probability that the designer holds each possible intention, thresholding intentions to act on. Treating intentions individually also gives the agent flexibility in how it acts on them. For example, the agent might choose to explore designs that optimize high-probability intentions or make less-binding suggestions optimizing for intentions near the probability threshold.

Ultimately, establishing joint intentions requires mutual awareness between collaborators that they share intentions, and a commitment to design for them together. Whatever the accuracy of the agent's model of what a human's intentions are, it must communicate this awareness. As such, a reasonable estimate of the probability that any design intention is held by a human teammate offers a basis for an agent to initiate communication about joint intentions, whether directly or implicitly through shared designs.

7.2 Toward a Better Probabilistic Model of Design Intentions

The usefulness of any model is dependent on its ability to simplify the real world without losing its ability to express the underlying phenomenon that it represents. Designing is an extraordinarily complex human process. We make several simplifying assumptions in our model to study the relationship between intentions and outcomes, but in doing so, we run the risk of losing aspects of complexity that are at the essence of what designing is.

To begin with, we chose not to model design space features or beliefs about the design space explicitly, focusing instead on design outcomes and intentions. However, both variables hold information that could influence how we predict design intentions. Firstly, certain design patterns may correlate with intentions without translating into good outcomes; accounting for this could increase the recall of our intention predictions. For example, many small districts in a population-dense region could indicate an intention to balance populations, even if outcomes do not yet reflect that.

Human Data Standardized Using	Class Accuracy			Label Precision				Label Recall				Label F1-Score			
	1st	2nd	3rd	Overall	P	V	C	Overall	P	V	C	Overall	P	V	C
Human Data	0.313	0.509	0.672	0.739	0.728	0.711	0.778	0.700	0.725	0.642	0.734	0.719	0.726	0.675	0.755
Agent Data	0.297	0.501	0.669	0.785	0.675	0.772	0.907	0.608	0.922	0.441	0.463	0.651	0.779	0.561	0.613

Table 2: This table compares the performance of our intentions model on human data that has been standardized using the distribution of all human design outcomes and the distribution of all agent design outcomes. This includes the class accuracy of the top one, two, and three predictions for each sample, as well as label-wise precision and recall for balancing population (P), improving voter efficiency (V), and making compact shapes (C).

Secondly, as a designer learns more about a design space, their design choices should more reliably produce good outcomes that align with their intentions. Removing the human's beliefs about the design space introduces an unexplained temporal dependency that weakens the somewhat tenuous Markov assumption that each new design is only dependent on the prior state, independent of the order of exploration. Our choice not to account for this dependency could impose a severe limit on our ability to fit outcomes to intentions. Other work models human beliefs in behavior [3], but accounting for and tracking human beliefs about the design space would significantly complicate our model.

Finally, while we model design intentions as static, they more likely evolve as designers explore a task. Designers may not actually know their intentions when they start designing, and what they learn about design dynamics and possibilities through exploration should change how they interpret the task. This widely held position has motivated cognitive and computational models of problem and solution *co-evolution* in designing [18, 19].

One reason we chose to model design intentions statically is the challenge of training a transition model for how humans tend to move from one intention to the next. Collecting human data for this task would require some means of probing human design intentions as they evolve, rather than pre-defining them as we did. This raises issues common to self-reporting, e.g. designers' difficulty expressing their intentions and the probe itself breaking the immersion of the design process. With that said, it may be possible to extract a transition model from agent-simulated behavior, as Maher *et al.* have demonstrated the feasibility of searching concurrently through design problem and solution spaces [18] and Grace and Maher have explored reframing design goals by modelling surprise [12].

7.3 Advantages and Pitfalls in Simulating Human Design Behavior

We envisage several potential benefits in simulating human design behavior with an agent. Clearly, simulated data can be more cost-efficient to collect at scale. Designing is cognitively demanding and typically consists of extended and thoughtful exploration, *after* developing a basic understanding of the task and design interface. Simulated design proxies can further be used to explore less popular or obvious subsets of the design space. Assigning intentions to participants, as we did, is only feasible with small intention sets, and assigning designers to intentions that they may not wish to adapt can potentially inhibit how naturally they explore solutions.

The idea of using agent behavior to infer human intentions is well-founded in theories of human cognition, in particular those ascribing to a variant of simulation theory [6, 29]. Similarly, our findings suggest that agents might be able to reason about human intentions based on their own experience.

Still, there is a clear gap in how our model generalizes from simulated training and validation data to the human data. This raises questions about what characteristics of agents affect their capacity to simulate human design behavior. Our ϵ-greedy proxy agent was simple and allowed for the likelihood that humans do not monotonically improve designs. However, non-optimal behavior may not be random choice–a human might choose a non-optimal design on a hunch that it will pay off in the long run, to learn more about the design space, or to test something they are uncertain about. Perhaps an agent designed to balance exploration and exploitation (e.g. using Bayesian optimization or reinforcement learning), could represent this kind of deliberate meandering more closely.

One way that we reconciled potential differences between the simulated design process and the real human design process was by standardizing the agent and human data each according to their own distributions. This step may not be fair or realistic; since it relies on information about the human data, it may be better to think of the model as predicting intentions from outcomes relative to some understanding of human behavior. Indeed, the distributions of raw outcomes under each intention suggest that human designers tended to find districts with more balanced populations, while the agents tended to waste fewer votes and draw more compact boundaries (Figure 8). If we standardize the human data to the underlying distribution of the agent's data instead, the model predictions reflect these shifts (Table 2). While the class-level performance measures are similar, the model tends to over-predict human intentions to balance population and under-predict intentions to minimize wasted votes or maximize compactness.

This underscores the difficulty of studying how proxy agents can more closely approximate human behavior. We can only speculate why the distributions are different, but there will always be limits to how well we can emulate the underlying processes that drive design exploration for a given task, and the differences can be highly contextual. One approach to handling this without losing scale could be to simulate using proxy agents that learn from humans, e.g. by taking into account human distributions from samples, or through learning by demonstration and policy shaping methods.

8 LIMITATIONS AND FUTURE WORK

One limitation of this work involves our choice of design task. Design tasks are notoriously contextual, and our findings and insights may not generalize to a breadth of tasks. In particular, we dealt with limited ill-definition, having only three possible design intentions for the human to choose from. Tasks with more, less explicit, and evolving design intentions should be explored in future work. At

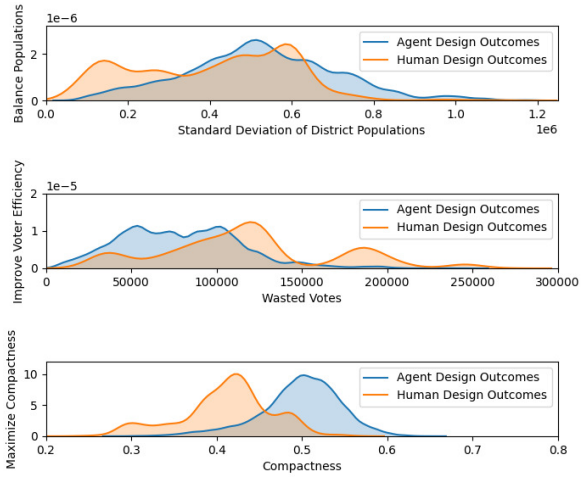

Figure 8: Density plots of relevant raw design outcomes for each intention, as explored by agent and human designers, suggest that, while humans found designs with more balanced populations, the agent tended to be more effective at improving voter efficiency and maximizing compactness.

the same time, we hope to incorporate contextual features about the design and the human designer's behavior into our predictions.

There are also limitations around our test set, which was generated by four members of the research team. Future work must evaluate whether these results generalize to a larger and more diverse test set, with considerations for important factors like designer expertise. A more robust set of human data could also be used to study how different features of search agents influence their effectiveness as proxies for human design behavior.

Finally, this work proposes a model to predict design intentions but does not integrate it into a collaborative design agent or evaluate how said predictions influence the ability of an agent to effectively collaborate with a human. Testing how intention-aware agents perform as teammates with human participants will be necessary to generate more practical design guidelines and evaluate the kinds of value that models of design intentions can provide.

9 CONCLUSION

In this paper, we studied intention inference as an important and difficult challenge en route to human-AI collaborative design through intelligent virtual agents. We propose a model to predict human design intentions from observed design outcomes and develop a data collection system to collect design behavior across multiple objectives. We fit our model on simulated design behavior and evaluate it on human test data. Our results suggest ways in which an agent's beliefs over individual intentions might prove useful in establishing joint intentions, as well as avenues for deeper study of using agents as data proxies for human designers.

10 ACKNOWLEDGMENTS

The authors thank Gonzalo Gonzalez-Pumariega for his contributions to the project. This work was supported primarily by the Civil, Mechanical and Manufacturing Innovation Program of the National Science Foundation under NSF Award No. 1907542.

REFERENCES

[1] Christopher Alexander. 1964. *Notes on the Synthesis of Form.* Vol. 5. Harvard University Press.
[2] James Allen, Nate Blaylock, and George Ferguson. 2002. A problem solving model for collaborative agents. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2.* 774–781.
[3] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.
[4] Amit Banerjee, Juan C Quiroz, and Sushil J Louis. 2008. A model of creative design using collaborative interactive genetic algorithms. In *Design Computing and Cognition'08.* Springer, 397–416.
[5] Hyunseung Bang, Antoni Virós Martin, Arnau Prat, and Daniel Selva. 2018. Daphne: An intelligent assistant for architecting earth observing satellite systems. In *2018 AIAA Information Systems-AIAA Infotech@ Aerospace.* 1366.
[6] Sarah-Jayne Blakemore and Jean Decety. 2001. From the perception of action to the understanding of intention. *Nature Reviews Neuroscience* 2, 8 (2001), 561–567.
[7] Michael E Bratman. 1990. What is intention. *Intentions in Communication* (1990), 15–31.
[8] Richard Buchanan. 1992. Wicked problems in design thinking. *Design Issues* 8, 2 (1992), 5–21.
[9] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems.* 4299–4307.
[10] Nicholas Davis, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An enactive model of creativity for computational collaboration and co-creation. In *Creativity in the Digital Age.* Springer, 109–133.
[11] John S Gero and Udo Kannengiesser. 2004. The situated function–behaviour–structure framework. *Design Studies* 25, 4 (2004), 373–391.
[12] Kazjon Grace and Mary Lou Maher. 2016. Surprise-Triggered Reformulation of Design Goals.. In *AAAI.* 3726–3732.
[13] Fazle Karim, Somshubra Majumdar, and Houshang Darabi. 2019. Insights into LSTM fully convolutional networks for time series classification. *IEEE Access* 7 (2019), 67718–67725.
[14] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. 2019. Multivariate lstm-fcns for time series classification. *Neural Networks* 116 (2019), 237–245.
[15] Nathaniel Kelso and Michal Migurski. 2017. Election Geodata. https://github.com/nvkelso/election-geodata.
[16] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980 (2015).
[17] Matthew V Law, JiHyun Jeong, Amritansh Kwatra, Malte F Jung, and Guy Hoffman. 2019. Negotiating the Creative Space in Human-Robot Collaborative Design. In *Proceedings of the 2019 on Designing Interactive Systems Conference.* ACM, 645–657.
[18] Mary Lou Maher and Josiah Poon. 1996. Modeling design exploration as co-evolution. *Computer-Aided Civil and Infrastructure Engineering* 11, 3 (1996), 195–209.
[19] Mary Lou Maher and Hsien-Hui Tang. 2003. Co-evolution as a computational and cognitive model of design. *Research in Engineering Design* 14, 1 (2003), 47–64.
[20] Soren Mindermann and Stuart Armstrong. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* Curran Associates Inc., 5603–5614.
[21] Andrew Y Ng, Stuart J Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.
[22] Don Norman. 2013. *The design of everyday things: Revised and expanded edition.* Basic books.
[23] Daniel D Polsby and Robert D Popper. 1991. The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale Law & Policy Review* 9, 2 (1991), 301–353.
[24] Horst WJ Rittel and Melvin M Webber. 1973. Dilemmas in a general theory of planning. *Policy sciences* 4, 2 (1973), 155–169.
[25] Paul A Rodgers, Avon P Huxor, and Nicholas HM Caldwell. 1999. Design support using distributed Web-based AI tools. *Research in Engineering Design* 11, 1 (1999), 31–44.
[26] Donald A Schon. 1992. Designing as reflective conversation with the materials of a design situation. *Research in Engineering Design* 3, 3 (1992), 131–147.
[27] Herbert A Simon. 2019. *The sciences of the artificial.* MIT press.
[28] Mohammad S Sorower. 2010. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis* 18 (2010), 1–25.
[29] David Vernon, Serge Thill, and Tom Ziemke. 2016. The role of intention in cognitive robotics. In *Toward Robotic Socially Believable Behaving Systems-Volume I.* Springer, 15–27.
[30] Willemien Visser. 2006. *The cognitive artifacts of designing.* CRC Press.
[31] Georgios N Yannakakis, Antonios Liapis, and Constantine Alexopoulos. 2014. Mixed-initiative co-creativity. In *FDG.*