An Information Bottleneck Characterization of the Understanding-Workload Tradeoff in Human-Centered Explainable AI

Lindsay Sanneman*
lindsays@csail.mit.edu
MIT
Cambridge, Massachusetts, USA

Mycal Tucker*
mycal@mit.edu
MIT
Cambridge, Massachusetts, USA

Julie A. Shah
julie_a_shah@csail.mit.edu
MIT
Cambridge, Massachusetts, USA

ABSTRACT

Recent advances in artificial intelligence (AI) have underscored the need for explainable AI (XAI) to support human understanding of AI systems. Consideration of human factors that impact explanation efficacy, such as mental workload and human understanding, is central to effective XAI design. Existing work in XAI has demonstrated a tradeoff between understanding and workload induced by different types of explanations. Explaining complex concepts through abstractions (hand-crafted groupings of related problem features) has been shown to effectively address and balance this workload-understanding tradeoff. In this work, we characterize the workload-understanding balance via the Information Bottleneck method: an information-theoretic approach which automatically generates abstractions that maximize informativeness and minimize complexity. In particular, we establish empirical connections between workload and complexity and between understanding and informativeness through human-subject experiments. This empirical link between human factors and information-theoretic concepts provides an important mathematical characterization of the workload-understanding tradeoff which enables user-tailored XAI design.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **User models**; **HCI theory, concepts and models**; **Empirical studies in HCI**; • **Mathematics of computing** → **Information theory**.

KEYWORDS

explainable AI, workload, human factors, information bottleneck

ACM Reference Format:
Lindsay Sanneman, Mycal Tucker, and Julie A. Shah. 2024. An Information Bottleneck Characterization of the Understanding-Workload Tradeoff in Human-Centered Explainable AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3630106.3659032

1 INTRODUCTION

With the rapid development of powerful yet opaque artificial intelligence systems, AI transparency methods that effectively explain the outputs of these systems to humans are increasingly important. Recent advances in explainable AI (XAI), defined as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future" [27], have aimed to address the problem of AI transparency. Accounting for human factors related to information processing is at the core of producing effective explanations that support human understanding of these AI systems. Such explanations require not only computer science expertise, but also cross-disciplinary efforts with the fields of cognitive science, human factors, and the social sciences generally [52, 53, 65].

Human-centered explainable AI and intelligibility research has begun to explore XAI with respect to human factors, both through proposed frameworks [3, 17, 19, 25, 45–47, 65, 76] and experimental or interview-based analyses [26, 41, 43, 55, 64, 87]. In particular, recent experiments have drawn upon validated assessments from human factors to study the impacts of XAI on human mental workload, trust, and conceptual understanding [34, 42, 43, 55, 64] and have demonstrated tradeoffs between these factors in explanation design [26, 43, 55, 64]. Such a tradeoff exists between increasing human workload and supporting human understanding of an AI system: one study found that explaining an autonomous agent's goals through hand-crafted abstractions of key problem features effectively balanced human workload and understanding as compared with other XAI techniques [64]. Further work has suggested that humans approach complex problem by using abstractions [32]; this suggests that providing abstract representations of key information may be an effective means of providing AI explanations to humans. However, generating such abstractions automatically and quantifying how they trade off human workload and understanding remain open problems.

Concurrent with such XAI and human factors research, recent cognitive science works have identified the key role of information-theoretic abstractions in human cognition. For example, in a wide variety of languages and semantic domains, human naming systems are near-optimal according to an Information Bottleneck (IB) tradeoff between maximizing informativeness (how well a listener can reconstruct a speaker's meaning) and minimizing complexity (how many bits about an input are encoded in a word) [54, 84–86]. Our key insight in this work is connecting concepts from IB literature to analogs in human factors for analyzing and designing explanations. By connecting notions from human factors, such as

*Authors contributed equally to this research.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659032

workload and understanding, to information-theoretic quantities, such as complexity and informativeness, we may leverage theoretical models for tradeoffs between these terms, while simultaneously using existing IB tools for automated explanation-generation.

In this work, we leverage IB to automatically generate abstract explanations and establish empirical connections between the aforementioned human factors constructs and information-theoretic quantities through a set of human-subject experiments. In particular, we consider the relationships between human workload and explanation complexity and between human understanding and distortion (a concept closely related to explanation informativeness that captures how well a feature can be predicted from a given abstraction). This lays the groundwork for modeling human factors quantities and their associated tradeoffs using the theoretically-rich computational IB framework, which will enhance our ability to automatically generate and analyze user-tailored explanations that account for differing human informational and workload needs. Here, we focus on explaining complex functions through abstractions of these functions, since many AI systems — including large language models (LLMs), reinforcement learning-based robotic systems, and machine learning-based recommender systems — are built on complex functions. Specifically, we study the problem of explaining reward functions (discussed at length in Section 2.2.1) to humans, with a focus on reward functions due to their applicability across many applications [68], such as autonomous agent planning problems [61, 71] and reinforcement learning from human feedback (RLHF), which is used to tune LLMs [9, 88].

We performed experiments in two domains — a grid-navigation domain and a color-based sample-collection domain — and considered both continuously and discontinuously varying reward functions. Our results indicate significant correlations between complexity and human workload as well as distortion and a feature-based measure of human understanding across both domains. We also observed significant correlations between a policy-based measure of human understanding and distortion within the grid-navigation domain, but not the color domain, which involved more complex visualizations of abstract explanations. These findings suggest that the complexity-distortion balance in IB can be effectively applied to model the workload-understanding tradeoff in human-centered XAI design and to generate user-tailored explanations, but that care must be taken in visualizing such abstract explanations.

2 RELATED WORK

Here we provide an overview of the literature on the human factors constructs of workload and human understanding as they apply to explainable AI as well as existing XAI approaches to explaining functions (including reward functions) to humans. We also discuss the Information Bottleneck method, which we leverage to automatically generate abstract explanations of reward functions. In our experiments, we hypothesize that the information-theoretic concepts used to generate these abstract explanations correlate with the human factors constructs of workload and understanding.

2.1 Human Factors and Explainable AI

2.1.1 Human Mental Workload. Mental workload is a widely-studied human factors construct that can be defined as the relationship between the mental resources demanded by a task and the resources available to be supplied by the human performing the task [56]. It has been researched in domains such as aviation [12, 39], healthcare [62, 74], and usability in human-computer interaction [48], among others [49], and has been shown to correlate with task performance across a variety of settings [14–16, 79]. Various models of mental workload have been proposed, such as the widely-applied multiple resource model (MRM) [77], which categorizes human cognitive resources into different independently-filled "pools" available for information processing. In the context of XAI, such models can inform workload considerations, such as only communicating information that is comprehensible to the explanation recipient. not process all necessary details. At the same time, explanations must provide adequate information in order to be useful for the task at hand. A person's available mental capacity, therefore, should inform the choice of the amount of information to present in an explanation [65]. Beyond this, accounting for individual differences in baseline cognitive capacity between people is critical to supporting human task performance [80, 81].

The impact of XAI on workload has been widely researched through both objective and subjective assessments, with some studies finding that the addition of AI transparency reduces workload due to increased access to critical information [17, 67, 83], some finding that additional information provided by XAI systems increases workload [26, 43, 55, 64], and others finding little impact of XAI on workload [14–16, 79]. Most relevant to our work, one study indicated that higher-complexity reward explanations were associate with higher workload and that providing abstract reward information mitigated increases in workload [64]. In this paper, we again consider abstract reward function explanations, and we study how workload relates to an information-theoretic concept of complexity in explanation design.

2.1.2 Human Understanding. In recent years, XAI researchers have explored assessments of explanation efficacy, including approaches that measure user comprehension/understanding of AI decision making processes [34, 42, 52]. Such measures include scales for explanation goodness [34, 42] and a user's ability to simulate an agent's optimal behavior [35, 42] (although often such measures have not been validated in human experiments). One construct from human factors, situation awareness (SA), provides a three-level framework for defining human contextual understanding through the identification of a human's informational needs given their role and context [21]. This framework has been applied to the development of transparency frameworks within the XAI literature [17, 65], and it has been operationalized through the validation of associated assessments of human situational understanding, such as the Situation Awareness Global Assessment Technique (SAGAT) [21, 22, 24]. Through the application of measures like SAGAT, SA has also been shown to correlate with task-related measures such as performance and error frequency [23]. Because of such validation, a SAGAT-like approach can be readily applied to measure understanding in the context of XAI. One recent study employed human-subject evaluations to validate reward alignment metrics that could be applied within assessments of contextual understanding like SAGAT in order to measure a person's understanding of a reward function [66]. We leverage a subset of these metrics in our analysis to study how a person's understanding relates to an information-theoretic concept of informativeness, and how this can be leveraged to design explanations that effectively trade off workload and understanding through the proxy measures of complexity and informativeness.

2.2 XAI Approaches to Explaining Functions

Many existing approaches to XAI strive to explain the often-complex functions that characterize AI models. For example, *feature importance techniques* explain the most important features in regression [50, 60], *saliency maps* unveil information about the gradients of the functions of neural networks and related models [1, 36, 51], *analogies* provide information about AI decisions by leveraging causal relationships from familiar common-knowledge domains [30, 31], and *rationalizations* summarize agent policies based on agent reward functions [18, 20], among others.

While some explanation approaches account for human workload by providing variable amounts of information in explanations depending on a user's cognitive capacity [13, 60, 69] and other works have discussed the importance of considering the impact of task complexity on human performance [11, 58, 63], to our knowledge, no other work has formally measured and accounted for the tradeoff between the human factors constructs of human workload and understanding in explanation design. Accounting for workload and understanding separate from performance is critical since there are many factors which contribute to user performance overall. In this paper, we empirically demonstrate the links between these constructs and mathematically-grounded information-theoretic concepts, which enables the automatic generation of explanations that trade off workload and understanding differently depending on user needs and capacities. We focus specifically on explanations of reward functions, which characterize desired autonomous agent behaviors in sequential decision-making problems such as reinforcement learning.

2.2.1 *XAI and Reward Functions.*

Reward functions are one of the primary components of Markov decision processes (MDPs), which are often used to model autonomous agent planning problems [71]. Within an MDP, the reward function characterizes the reward an agent receives for taking different actions from different states; in other words, reward functions dictate what optimal agent behavior (often referred to as the agent's policy) will look like within a given domain. Reward functions are often defined as follows:

$$R(s) = \omega^T \Phi(s). \tag{1}$$

Here, $\Phi(s)$ is a set of features whose values can be calculated based on the agent's state in the world (s), and ω is a set of weights indicating the trade-offs between these features.

Within existing XAI literature, reward functions have been explained through means including policy summaries which demonstrate roll-outs of optimal agent behavior originating from a variety of world states [4–6, 29], language-based rationalizations of agent policies [18, 20], techniques that reconcile a human's reward function with that of an agent [72], counterfactual demonstration-based explanations of key reward features [44], and decompositions of interpretable reward components provided to human users [7, 38], among others. One recent study found that explaining reward functions through abstractions of reward features effectively balanced a workload-understanding tradeoff among different reward explanations [64]. Since the abstraction-based approach proved effective in that study, we also evaluate abstract explanations of reward features in this work.

2.3 Information Bottleneck

We leverage methods from Information Bottleneck (IB) literature to formalize a tradeoff between complexity and reward distortion, which we then connect to human factors. In canonical IB settings, one seeks to generate (lossy) representations, Z, of inputs, X, which are used to predict a downstream quantity, Y [2, 73]. In this work, we only consider predicting a reward, Y, from features, X, but the IB framework is more widely applicable. The IB maximization problem is formulated as a tradeoff between two information-theoretic terms:

$$\text{maximize} \quad I(Y;Z) - \beta I(X;Z) \tag{2}$$

where β is a scalar parameter, $I(Y;Z)$ is the informativeness (measured as the number of bits about the reward Y retained in Z), and $I(X;Z)$ is the complexity (measured as the number of bits about the features X in Z). The IB formulation seeks to maximize informativeness while minimizing complexity. Notably, there is a theoretical limit for the maximum informativeness for a given complexity, but this limit shifts as a function of β. In our work explaining reward functions, X is the features of the reward function, $\Phi(s)$, Y is the reward value, and Z are the abstract representations grouping X to predict Y. Therefore, as β increases and complexity decreases, the above optimization will group features with similar rewards in the same abstraction. Lastly, we note that IB work is closely related to rate distortion theory where, rather than computing informativeness ($I(Y;Z)$), one measures the distortion, or error, in predicting Y from Z [84]. In our work, we measure the distortion in predicting a reward value, which we dub *reward distortion*.

Beyond a purely mathematical formulation, several works in the fields of cognitive science, psychology, and behavioral economics have investigated aspects of IB tradeoffs in human cognition. Across domains and languages, naming systems (e.g., words for colors, family relatives, pronouns, etc.) are nearly perfectly efficient in the IB sense: maximizing the ability of listeners to reconstruct a speaker's meaning at a given complexity level [54, 84–86]. In vision-based domains, people similarly create compressed representations of images via sketches that capture functionally useful details at the expense of visual fidelity [37]; this type of behavior is consistent with an IB system under complexity constraints. In economics, recent research points to the importance of information constraints in human behavior [8]. Even within XAI, Bang et al. [10] briefly explored the role of penalizing complexity to create more "interpretable" AI models for humans to understand, but they only used a fixed complexity in experiments. This evidence, collected across multiple fields, suggests that IB tradeoffs play an important role in human cognition; in our work, we connect notions from IB to human factors measures of explanation understanding.

3 RESEARCH AIMS

In this paper, we aim to establish an empirical link between human factors concepts relevant to the design of effective explainable AI systems and information-theoretic concepts. This enables a mathematical characterization of tradeoffs between relevant human factors in XAI design and provides tools to automatically generate abstraction-based explanations which trade off these factors, which is useful for meeting the varying informational and workload needs of individual users of AI systems. Specifically, we perform human-subject experiments in order to validate an information-theoretic measure of explanation complexity as an indicator of human workload and an information-theoretic measure of reward distortion as an indicator of human understanding. Next, we detail the human factors concepts and information-theoretic metrics we studied in greater detail, along with the hypotheses assessed in our experiments.

3.1 Measures of Human Workload and Human Understanding from the Field of Human Factors

3.1.1 Workload. As discussed in Section 2.1.1, multiple measures of human workload have been applied within the field of human factors. One of the measures most commonly applied within the literature is the NASA Task Load Index (TLX) scale [28], which asks respondents to answer a set of Likert scale-based questions about their workload after completing a task. NASA TLX has also been applied to the study of various techniques for explainable AI [26, 43, 55, 64, 82]. We therefore used the NASA TLX scale to assess workload in our own set of experiments.

3.1.2 Human Understanding. While a number of approaches have been proposed for assessing human understanding in the context of XAI [34, 35, 42], as discussed in Section 2.1.2, few have been validated through human-subject experiments. One recent set of experiments validated reward alignment metrics which capture the similarity between a human's reward function and that of an autonomous agent [66]; these metrics can be applied to study either how aligned an agent's reward function is with a human's after a reward-learning process on the one hand, or how aligned a human's understanding of an agent's reward function is with the agent's true reward after an explanation of the reward is provided on the other. As our aim is to study human understanding of reward functions resulting from the provision of reward explanations at different levels of abstraction, we consider the latter application in this work. Note that we consider scenarios in which reward information constitutes part of the human's overall SA, and assessing reward understanding is therefore a crucial component of an overall SAGAT-based analysis, as discussed in Section 2.1.2.

Sanneman and Shah [66] identified two categories of alignment in their experiments: feature alignment, which captures how aligned human and agent reward features and weights are; and policy alignment, which captures how aligned human and agent policies corresponding to these reward functions are. We leverage one of the validated metrics from each of these categories in our assessments of human understanding.

We apply a similarity metric called *feature ranking* to evaluate feature understanding, defined as follows:

$$FR = \frac{(W_H \cap W_{GT})}{(W_H \cup W_{GT})} \tag{3}$$

Here, W_{GT} is the set of pairwise comparisons of the magnitudes of the weights w of a set of reward features, $\Phi(s)$, as in Equation 1 (e.g., one of these comparisons could be $w_A > w_B$, where w_A is the weight of feature A, which is higher than w_B, the weight of feature B). W_{GT} specifically captures the pairwise rankings of the feature weights in the ground truth reward function, and W_H is the set of pairwise rankings from the human's reward function. In our experiments, we assessed the human's reward function by asking the human participants to rank a set of features in order of importance according to their understanding of the reward function upon receiving an abstract explanation of this reward. (We include examples of this assessment in Appendix D.) The *feature ranking* metric is the intersection over union of the ground truth rankings and the human's rankings, and thus captures the similarity between how important the human believes a set of reward features are relative to each other versus the ground truth relative importance of each feature.

The policy understanding metric we apply is a regret-based metric called *best demonstration*, defined as follows:

$$BD = 1 - \frac{R(\xi^*) - R(\xi^H)}{R(\xi^*) - R(\xi^-)} \tag{4}$$

Here, ξ^* is the optimal demonstration (i.e., a set of state-actions pairs) of a given task according to a ground truth reward function — which, in our case, is the reward function being explained. ξ^H is the human's best demonstration according to the reward function they understood from the explanation. Finally, ξ^- is the worst possible demonstration in terms of the ground truth reward, which we assume to be calculable for a finite-horizon task. $R(\cdot)$ evaluates each of these trajectories according to the ground truth reward function. To evaluate this metric in our experiments, we asked human participants to provide an optimal demonstration of a task given their understanding of the explained reward function (examples depicted in Appendix D). The *best demonstration* metric essentially captures how close the human's reward function is to the ground truth reward in terms of the policies that result from these rewards for a given task.

3.2 Information-Theoretic Measures of Explanation Complexity and Reward Distortion

3.2.1 Complexity. Drawing upon prior literature, we define *complexity* as the mutual information between an input, X, and an abstraction, Z [75, 84]. This measure is defined via the Kullback–Leibler divergence of the conditional distribution of Z given X from the prior over Z:

$$I(X;Z) = D_{\mathrm{KL}}[\mathbb{P}(Z|X) \| \mathbb{P}(Z)]. \tag{5}$$

Complexity is minimized at 0 if all X are represented via the same Z; beyond such uninformative representations, more complex representations include additional information about X in Z. For example, more complex color naming systems use more distinct words: naming systems using the word "crimson" convey more information about a precise color than naming systems that only use less specific words like "red" [84]. More generally, in IB literature, by penalizing the complexity of representations, one imposes a bottleneck on how much information is stored in Z, which in turn induces lossy representations that do not enable perfect reconstructions of X from Z [2, 73]. In our work, we both use existing IB methods to generate representations across a spectrum of complexity, as well as calculate complexity as a metric which we then relate to human workload.

3.2.2 Reward Distortion. We define reward distortion as a measure for how well one can predict a reward value, Y, from an abstract representation, Z. Formally, we measure reward distortion as the minimum mean squared error (MSE) for predictions of Y from Z:

$$D(Z;Y) = \frac{1}{|Y|} \sum_{(X,Y)} ||\hat{Y}(Z(X)) - Y||^2 \tag{6}$$

where, assuming access to a dataset of reward features (X) and rewards (Y), $Z(X)$ represents the abstraction generated from X, and $\hat{Y}(Z(X))$ represents the optimal prediction of Y given $Z(X)$. With a small set of discrete representations, Z, computing an optimal predictor is equivalent to traditional methods for MSE regression. We note that reward distortion is similar to notions of informativeness ($I(Y;Z)$) from traditional IB literature, and rate distortion theory directly considers tradeoffs between distortion and complexity (e.g., see Zaslavsky et al. [84]). Given the continuous nature of reward values, we therefore use reward distortion as our preferred metric. In our paper, we seek to connect reward distortion to *feature rank (FR)* and *best demonstration (BD)* metrics of human understanding. Given that we always measure the distortion in predicting reward, we at times refer to *reward distortion* simply as *distortion.*

3.3 Hypotheses

We investigated the impact of varying the complexity and distortion of abstract explanations of reward functions on human workload and human reward understanding by evaluating the following hypotheses:

Hypothesis 1. *The distortion of the explanations will be negatively correlated with human reward understanding, including both feature and policy understanding.*

Hypothesis 2. *The complexity of the explanations will be positively correlated with human mental workload.*

Jointly, these hypotheses state that decreasing distortion will improve human understanding (H1), but increasing the complexity of abstractions will result in greater workload (H2). In other words, we aim to evaluate whether (the inverse of) distortion is a suitable proxy for human understanding and whether complexity is a suitable proxy for human workload in explanation design. Given theoretical bounds from IB literature showing a minimum distortion for a given complexity, this suggests optimal explanations will also trade off these two competing factors: minimizing distortion to achieve a desired level of understanding, subject to bounds on workload (and therefore complexity).

4 METHODOLOGY

4.1 Domains

We leveraged two domains in our experiments: a grid-based navigation domain and a color domain. Participants could optimally solve associated tasks across both (determined through pilot studies), eliminating a "solve-ability" confound.

4.1.1 Grid Navigation Domain. In the grid-navigation domain, different reward values between -1 and +1 were assigned to squares in a 5x5 grid, as depicted in Figure 1. The task was to navigate between a start square and goal square while maximizing the reward accumulated along the path. In this set of experiments, we considered two different reward functions: the first, the "Manhattan grid" depicted in Figure 1 (a), had a maximal reward value of +1 at one of the grid squares, with the reward values of other squares in the grid decreasing according to the Manhattan distance from that square. In the second, the "random grid" depicted in Figure 1 (b), the reward for each location was sampled uniformly at random within the range $[-1, 1]$. We studied these two reward functions as examples of: (1) a continuously varying reward function where adjacent grid regions have similar reward values (the Manhattan grid); and (2) a discontinuously varying reward function, where adjacent grid regions do not necessarily form natural abstraction groups (the random grid). All abstract representations of the grid regions were provided as heat maps, as shown in Appendix C.

4.1.2 Color Domain. In the color domain, we applied continuous and discontinuous reward functions to the colors depicted in Figure 1 c. We drew the colors from the World Colorchip Survey (WCS) dataset [40]. Notably, prior literature has identified how languages represent colors at different abstraction levels, which motivated our study of abstraction-based explanations for improving human understanding of color-based reward functions within this domain [84]. For the continuous reward function, we set each color's reward equal to that of the blue value in the RGB representation (between 0 and 1). For the discontinuous reward function, we used a hand-specified function that divided the blue values into eight bins, which we assigned different rewards between -1 and 1; the exact function is in Appendix A.2. Thus, the color domain largely mirrored the grid navigation domain by establishing both continuous and discontinuous reward functions. The task that participants performed for the *best demonstration (BD)* assessment in the color domain was a sample collection task, where the objective was to navigate through a grid and maximize the total value of samples collected along the path. The samples were represented by one of the colors in the original color grid, and for consistency with the grid domain, abstract representations of the color regions were provided as heat maps (Appendix C).

4.2 Information-Bottleneck Explanation Generation

We used existing methods to generate explanations of our reward functions at different complexity and distortion levels. We used an existing IB solver (embo), which only require as inputs a joint distribution of inputs and rewards: $\mathbb{P}(X, Y)$ [57, 70]. For each domain, therefore, we computed this joint distribution by iterating over all possible inputs X (e.g., the (x, y) location of a cell in the

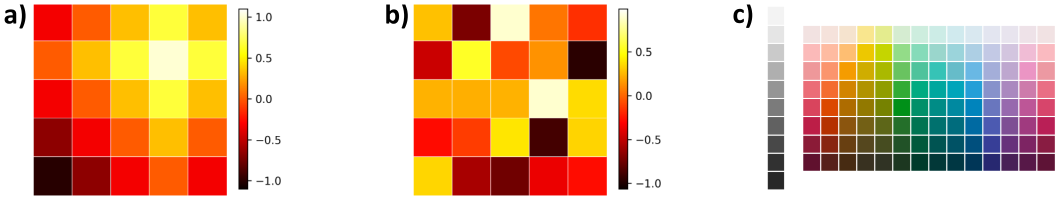


Figure 1: Our three experiment domains. In the Manhattan (a) and Random (b) grids, reward was Manhattan distance from a fixed point or randomly-distributed. In the color domain (c), reward was based on the blue value in a color's RGB representation.

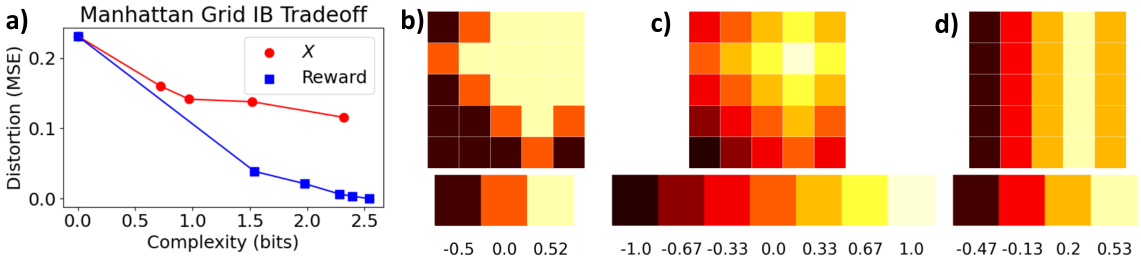


Figure 2: Complexity-distortion curves (a) and corresponding abstractions (b-d) for the Manhattan grid navigation domain. Using the true grid reward leads to low distortion as complexity increases (blue "Reward" curve), and more fine-grained abstractions (b-c), eventually recovering the underlying reward grid. Generating abstractions to recover the x coordinate in the grid (red "X" curve and d), rather than the reward, led to higher distortion due to abstractions that did not align with the true reward structure.

grid world) and computing the associated reward Y (e.g., the value at that location); to apply our method to other domains, one would need to provide a different $\mathbb{P}(X, Y)$ describing inputs and rewards. Further details of this process are included in Appendix A.

For a given reward function, the IB method generated abstractions for different solutions trading off distortion and complexity; we dubbed such abstractions the "reward-optimal" abstractions. Increasing the complexity of reward-optimal abstractions led to more fine-grained abstractions and lower distortion, as shown in Figures 2 b (low complexity) and c (high complexity). In our experiments, however, we wished to explore the effects of varying distortion and complexity independently. Therefore, in addition to the reward-optimal abstractions, we generated additional abstractions using different "training objectives:" alternate reward functions, which were not necessarily relevant to the structure of the reward function being explained. The abstractions generated by these alternate training objectives resulted in higher distortion (with respect to the true reward function) for the same complexity as the reward-optimal abstractions.

For example, in the Manhattan grid, one training objective we used was predicting the x location of each cell as the reward value. Abstractions generated from this function represented vertical strips in the grid, as shown in Figure 2 d. At the same time, we evaluated the distortion of such abstractions by measuring the MSE in predicting the actual reward in the Manhattan grid. (Informativeness and distortion are tightly linked concepts, as explored in information theory; we use distortion as an evaluation metric because we believe it corresponds to more intuitive notions of the quality of an explanation.) Figure 2 a shows how, in general, using x-based abstractions led to higher distortion, for the same complexity, than using reward-optimal abstractions based on the true Manhattan reward.

The same trends held in the color domains as well, where we generated reward-optimal abstractions based on the continuous or discontinuous reward functions of a color's blue value (from its RGB representation), as discussed in Section 4.1.2. To generate non-reward-optimal baseline abstractions in these cases, we used the color's red value (again, from RGB) as an alternate training objective; such abstractions were not useful in predicting the true reward value (which depended only on the blue value), leading to high distortion regardless of complexity. (See Figures 6 for complexity-distortion curves in the color domain.) Overall, by using a variety of training objectives to generate abstractions, we could test explanations using abstractions at the same complexity, but different distortion, levels. Appendix C includes examples of abstractions at different complexity levels and training objectives, in all domains.

4.3 Experiment Design

In order to empirically study the relationships between abstraction complexity and human workload and between distortion and human understanding, we performed human-subject experiments across both domains. In each, we studied explanations of the two types of reward function (continuous and discontinuous) discussed in Section 4.1. For each reward type, we generated a set of abstract explanations spanning a range of complexities and distortions. (We include examples of these abstractions in Appendix C.) The independent variables were the complexities and distortions of the abstractions; dependent variables included both human mental workload (as measured via the NASA TLX scale) and reward understanding, measured according to *feature ranking (FR)* and *best demonstration (BD)* assessments.

4.4 Procedure

For each of the two experiments (one in each domain), participants were first asked a set of demographic questions, including a question about colorblindness, since successful performance of the provided tasks relied upon interpretation of color. Next, participants received an overview of the experiment. In order to reduce learning effects and ensure that participants understood the task, they were presented with a set of example abstract explanations, along with corresponding examples of correct responses to the two reward understanding questions that they were asked throughout the experiment (*feature rank* and *best demonstration* as detailed in Section 3.1.2). Following these examples, participants were presented with two different scenarios in the given domain. For each scenario in each experiment, the participant was shown abstractions based upon a combination of training objective and complexity level.

Grid Domain: In the grid domain, one scenario involved the Manhattan grid (based on a continuously varying distance from a point), and one scenario involved the random grid (a discontinuous reward function). Abstractions were selected from the set of three training objectives (continuous blue, discontinuous blue, and red) and five possible complexities, for a total of 15 possible abstractions. The five complexity values were chosen such that the abstract colors regions spanned a range from one to eight regions.

Color Domain: In the color domain, one scenario involved the continuous reward function, while the other involved the discontinuous reward function. The order of presentation of scenarios was counterbalanced for each experiment across all participants. As in the grid domain, there were three training objectives (Manhattan or random reward, y-based reward, and x-based reward) with five possible levels of complexity, for a total of 15 possible abstractions. The number of abstract color regions again spanned a range of one to eight.

Following each scenario, participants were asked the feature rank and best demonstration questions; they were also asked the six NASA TLX scale questions to assess cognitive workload, along with seven additional questions related to subjective assessment of the explanation quality, which were adapted team fluency scale questions from Hoffman [33]. At the end of the experiment, participants were asked open-ended feedback questions about the experiment, including what they found to be most challenging and whether they had feedback about the experiment. In addition, we asked two attention-check questions during the survey: one before the two scenarios, and another immediately after.

4.4.1 *Experiment Details and Participants.*

We administered both experiments through the Qualtrics platform and recruited our participants via the Prolific platform. Participants received no time limit and took an average of 28 minutes to complete the color survey and 19.5 minutes to complete the grid navigation survey. They were compensated with $7 for completing the grid navigation survey and $10 for completing the color survey, with bonus payments of $20 provided to the highest-performing participants in each case. As this experiment involved minimal risk, it qualified for exempt human-subject evaluation status according to the policies outlined by MIT's institutional review board (IRB).

5 RESULTS

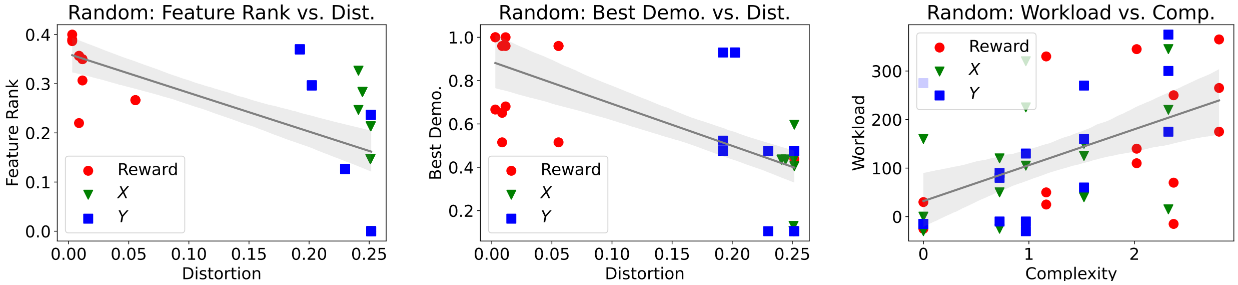
5.1 Grid Navigation Domain

Fifty-one participants completed the grid navigation survey (20 women, 30 men, and 1 non-binary individual). The median age was 37 years (min=19 years, max=76 years). Data from six participants was omitted from analysis due to failed attention-check questions or incomplete responses. The number of survey participants was chosen in order to ensure that we recorded at least three responses for each of the 15 different possible abstraction level-training objective pairs (five levels of abstraction and three training objectives). We first analyzed the Spearman correlations between distortion and understanding and between complexity and workload for the each of the Random and Manhattan grids separately. We used Spearman correlations due to the non-normality of the underlying datasets in this analysis, as well as the expected monotonic relationships between the correlated variables. We then analyzed the combined Random and Manhattan grid data through a linear mixed-effects analysis to account for individual differences in participants' responses to the reward understanding and workload questions.

Results from the Random grid domain are in Figure 3. Both feature rank (FR) and best demonstration (BD) were negatively correlated with reward distortion (Figures 3 a and b). Intuitively, this supports our hypothesis that understanding would decrease as reward distortion increased (H1). Quantitatively, these results were significant: using the Spearman correlation coefficient, we found $\rho(\text{FR}, \text{Dist}) = -0.83, \quad (p < 0.001)$ and $\rho(\text{BD}, \text{Dist}) = -0.68, \quad (p < 0.001)$. At the same time, Figure 3 c shows a significantly positive correlation between workload and complexity ($\rho(\text{Work., Comp.}) = 0.53, \quad (p < 0.001)$); that is, as the complexity of abstractions increased, so did the workload, supporting H2.

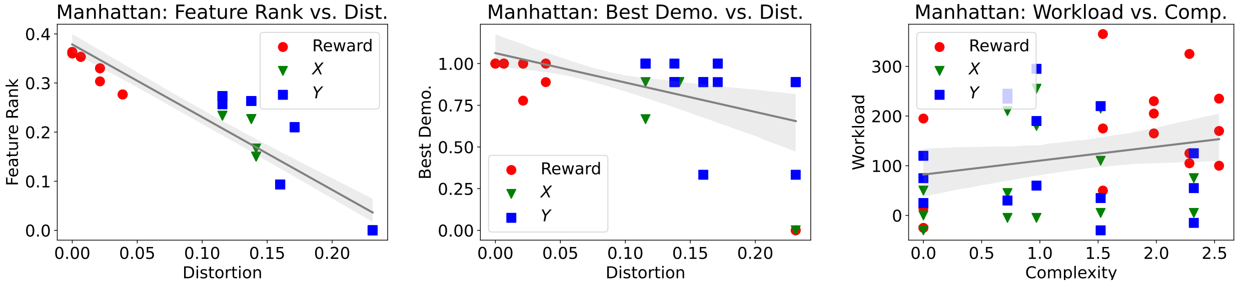
We observed similar trends in the Manhattan grid domain, depicted in Figure 4. Both FR and BD scores decreased as distortion increased (Figures 4 a and b). Concretely, $\rho(\text{ FR, dist}) = -0.97 \ (p < 0.001)$, and $\rho(\text{ BD, dist}) = -0.63 \ (p < 0.001)$. That is, each measure of understanding was significantly negatively correlated with distortion and, as before, the correlation was stronger for FR than for BD. Unlike with the Random grid, we did not identify a significant correlation between workload and complexity, although the trend was still positive: $\rho(\text{ Work., Comp.}) = 0.23 \ \ p = 0.13$.

Following the analysis of Spearman correlations for each grid type separately, we performed linear mixed effects modeling (LMEM) on the joint data from the Manhattan and Random grids and found significant trends supporting all our hypotheses. The models we applied for this analysis were formulated according to the following equations in Wilkinson notation [78]: $FR \sim Distortion + (1|Participant)$, $BD \sim Distortion{+}(1|Participant)$, and $Workload \sim Complexity + (1|Participant)$. Here, the models were fit using $Participant$ as a grouping variable, with a random intercept to account for the individual differences between participants, which were not accounted for by the Spearman correlations (e.g., perhaps one participant would consistently report higher workload). In our joint analysis, we leveraged the fact that each participant answered



(a) Feature Rank vs. Distortion **(b) Best Demonstration vs. Distortion** **(c) Workload vs. Complexity**

Figure 3: Results from the Random grid domain. As distortion increased, explanation understanding, as measured by Feature Rank (a) or Best Demonstration (b), decreased. At the same time, as complexity increased, workload increased (c).



(a) Feature Rank vs. Distortion **(b) Best Demonstration vs. Distortion** **(c) Workload vs. Complexity**

Figure 4: Manhattan domain results, connecting metrics of understanding to distortion (a-b) and workload to complexity (c). Increased distortion led to worsened understanding, and increased complexity led to increased workload.

one question about the Manhattan grid and one about the Random grid. We observed significant main effects for distortion and complexity within each model at the $p < 0.001$ level, with the following effect sizes and intercepts for each: $FR = -0.92 * \text{Dist.} + 0.35$, $BD = -2.19 * \text{Dist.} + 1.02$, $\text{Work.} = 33.50 * \text{Comp.} + 79.80$.

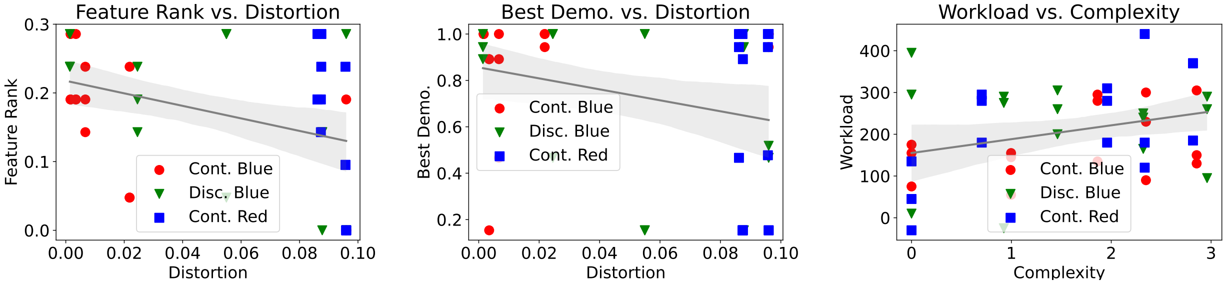
Overall, our grid domain experiment results strongly support our hypotheses that 1) decreased distortion would be associated with increased understanding and 2) increased complexity would be associated with increased workload. We found statistically significant support for all but one of our hypotheses via Spearman correlation tests assessing each grid-based reward function separately, and we found support for all of our hypotheses when evaluating the combined datasets through a linear mixed effects analysis, which accounted for individual differences in participant responses.

5.2 Color Domain

As in the grid domains, in the color domain we analyzed the results of the continuous and discontinuous reward functions separately via Spearman correlation and then performed joint analysis of the full data with a LMEM. Forty-seven participants completed the color survey (11 women, 33 men, 1 non-binary individual, and 2 not reporting gender). The median age was 33 years (min=20 years, max=66 years). We omitted data collected from two participants from analysis due to failed attention-check questions or incomplete responses. The number of participants was selected in the same manner as in the grid domain.

Results from the color domain experiments corroborate the key trends observed in the grid navigation domains, with the exception of the *best demonstration (BD)* understanding assessment. First, separating results by continuous and discontinuous reward functions, we established significant Spearman correlations for some of the hypothesized trends. *Feature rank (FR)* and distortion were negatively correlated for both reward functions ($p < 0.001$). For the continuous reward (shown in Figure 5 a), $\rho(\text{FR}, \text{dist}) = -0.47$; for discontinuous, $\rho(\text{FR}, \text{dist}) = -0.40$. Correlations between complexity and workload were positive, but not at the $p = 0.05$ level: for continuous (Figure 5 c), $\rho(\text{Work.}, \text{Comp.}) = 0.24$ $(p = 0.06)$ and for discontinuous, $\rho(\text{Work.}, \text{Comp.}) = 0.18$ $(p = 0.11)$. Lastly, correlations between the *best demonstration (BD)* understanding assessment and distortion were weak, with no significance value lower than 0.15. We attribute the *best demonstration (BD)* correlation failure to high random chance performance with high distortion: even with completely uninformative abstractions, some participants selected the optimal best path through random guessing. Also of note is that the visualizations of the abstractions within the color domain were not provided in the same space in which the best demonstration task was performed (as opposed to the grid domain, where the abstractions were visualized in the grid itself), so another possible reason for this difference (and the added difficulty with high-complexity abstractions) is the extra step necessary to translate the abstract information into the task space.

Complementing our Spearman correlation analysis for each domain separately, we again performed LMEM tests on all the joint data, grouped by participant, as we did for our grid navigation



(a) Feature Rank vs. Distortion **(b) Best Demonstration vs. Distortion** **(c) Workload vs. Complexity**

Figure 5: Color domain results, using the continuous reward function. Trends were weaker than those observed in the grid domains, but still reflected the hypothesized directions. Similar results for the discontinuous reward function are included in Figure 7 in Appendix B.

experiments. We applied linear mixed effects models of the same form, and found significant main effects for both distortion in the *FR* model and complexity in the *workload* model. The linear trends for *FR* vs. distortion and workload vs. complexity were significant at $p < 0.05$: $\text{FR} = -0.70*\text{Dist}+0.34$, $\text{Work.} = 39.43*\text{Comp.}+156.23$. These findings support our two hypotheses that (H1) increasing distortion would decrease understanding and (H2) increasing complexity would increase workload. As before, however, the correlation between *BD* and distortion was not statistically significant: $\text{BD} = -1.11 * \text{Dist.} + 0.73 \ \ (p = 0.12)$.

6 DISCUSSION

Across domains, we found evidence supporting our two hypotheses. In every domain, and with every reward function, we observed significantly negative Spearman correlation coefficients between feature-based understanding and distortion. While the significance of other trends varied slightly across domains, when we leveraged the within-participant aspect of our experiment design to account for individual differences in participant responses, we similarly found significantly positive correlations between workload and complexity. Although weaker than the *FR* and distortion correlations, we also found significant correlations between the *BD* measure of understanding (i.e., policy understanding) and distortion in all cases within the grid navigation domain. These weaker correlations track with previous experimental results, which demonstrated that the factor loadings for policy-based assessments of alignment (understanding) were weaker than those for feature-based assessments [66]; this is likely due to the additional challenge of translating a reward function into an optimal policy within a given environment. Nonetheless, the overall trends in our results support the hypothesized link between human factors constructs and information-theoretic concepts. This enables us to leverage these information-theoretic concepts to mathematically characterize the workload-understanding tradeoff in XAI design and to automatically generate abstraction-based explanations which trade off these factors, allowing us to account for variable informational and workload needs between different users of AI systems.

Overall, we observed a larger number of significant results and stronger correlations in the grid-navigation domain compared with the color domain, particularly for the *BD* (policy understanding) results. This is likely related to the visualizations of the abstractions in each domain: in the grid-navigation domain, heat maps of square values were provided within the best demonstration task grid itself, while in the color domain, participants interpreted the heat maps and their relation the color grid separately from the task grid, and then had to translate their reward understanding into an optimal policy in the task grid in an additional step. Beyond this, identifying the precise differences between individual colors within the color grid may have posed an additional challenge. While we identified significant support for our key hypotheses related to the relationships between information-theoretic concepts and human factors constructs in abstract explanation generation across both experiments, the differences in the *BD* results corresponding to the different types of abstraction visualizations between the experiments highlight the importance of carefully considering how to visualize abstractions for effective communication in future explanation design.

Finally, we observed some evidence that continuous reward functions may be better candidates for abstraction than discontinuous ones. The correlations between *FR* (feature-based reward understanding) and distortion were stronger for the Manhattan grid (with a fundamentally continuous reward structure) than the Random grid in the grid-navigation domain, and for the continuously varying reward function than the discontinuously varying reward function in the color domain. This suggests that abstracting such continuously varying reward regions may lead to more natural explanations of reward functions than groupings of discontinuous reward regions. We leave additional exploration of the impact of the structure of reward functions on explanation efficacy as an area for future work.

6.1 Limitations and Future Work

Our work takes an important step toward connecting IB methods to human factors in understanding explanations, but we rely upon some simplifying assumptions. First, in our experiments, we used the "ground truth" reward function, as well as alternative reward function baselines, to generate abstractions. As shown in our results, generating low-distortion explanations is extremely important for understanding explanations; future work may wish to further consider methods for generating good abstractions without access to reward functions. Second, we find support for linear relationships between the human factors and information theoretic concepts explored in our experiments through linear-mixed effects analysis. However, non-linear relationships might exist, particularly between workload and complexity, where reducing the complexity of explanations beyond some point might ultimately increase the difficulty of interpretation and correspond to an increase workload. Third, we used exact IB methods for generating abstractions, which may not scale to larger domains. Fortunately, recent work proposed approximate discrete IB methods at large scale [75]; we anticipate that such methods may be easily combined with our explanation work.

Finally, we scoped our work to study abstract explanations of reward functions in particular, but there are additional concepts related to AI decision making, such as policies, constraints, and decision uncertainty (among others), which might also be effectively explained through automatically generated abstract explanations. Beyond this, our technique can be extended to explain complex concepts in larger-scale domains, such as reinforcement learning-based robotics applications and large language models (LLMs). In particular, given recent emphasis on reinforcement learning from human feedback (RLHF) and connections between rewards and policies (as in DPO [59]), we believe that lossy, simplified explanations of complex reward functions for language models are a promising direction for future work. In this work, we have established empirical connections between information-theoretic concepts and human factors constructs which we hope will apply to explanation design for this broader scope of AI concepts and domains, and have laid the groundwork for future exploration and confirmation of these relationships in different settings.

7 CONCLUSION

In this work, we established empirical connections between human factors metrics of explanation understanding and workload with Information Bottleneck (IB) concepts of distortion and complexity. In the standard IB framework, a tradeoff exists between distortion and complexity; we established a similar tradeoff of people improving reward understanding as distortion decreased, but at at the cost of increased workload as complexity increased. Our findings may be used directly to inform explanation design, especially in accounting for differing informational and workload needs between individual users of AI systems. For example, given a maximum acceptable workload, one could find the corresponding allowed complexity level for explanations, and, at that complexity level, promote understanding by minimizing distortion. More generally, our work establishes important connections at the intersection of human factors and information theory, which we hope future work will continue to explore.

ACKNOWLEDGMENTS

This work was supported by the Army Research Laboratory (ARL) under Grant ARL DCIST CRA W911NF-17-2-0181.

8 ETHICS, UNINTENDED IMPACTS, AND AUTHOR BACKGROUND

Our work is inspired by a fundamentally ethical view of human-centered explanation design, but there remains the possibility of unintended consequences related to mis-aligned representations and interpretations. On the one hand, our work seeks to establish important theoretical and experimental guidance on how to best design AI explanations to support human understanding while limiting workload. Such guidance should support better user experiences as well as personalization (e.g., allowing individuals to specify their own workload/understanding tradeoff preferences). On the other hand, the real world may not be as supportive of allowing individual variation across the tradeoff spectrum we explore. For example, different stakeholder incentives may lead to engineers or designers opting for high-understanding/high-workload explanations that actual users who have to interpret explanations find more challenging. Furthermore, as demonstrated in experiments, the choice of reward function in generating explanations has important effects on explanation understandability. In all our experiments, we assumed that there was a fixed "ground truth" reward function; if different users prioritize different features or behaviors (i.e., have different internal reward functions), our IB tradeoff framework alone is insufficient to address such varied needs. Thus, while we begin to explore aspects of personalization in explanations, societal risks, in particular revolving around who controls the personalization, remain.

Beyond broad ethical considerations, we authors recognize the role that our backgrounds, educations, and experiences play in developing this work. We hold that human-centered post-hoc explanation design, using abstractions that may come at the expense of providing the most faithful explanation, is an important area of research. Other researchers can and do argue that approximate post-hoc explanations should be avoided. This debate is clearly important, but given the prevalence of black-box systems with no explanation mechanisms at all, we believe our work still holds practical value. Our backgrounds in the fields of human factors, computer science, and cognitive science biases us towards thinking about lossy cognitive models. We believe that mathematical models of human cognition and explanations offer important insight into how to better design human-compatible technologies; other researchers with different backgrounds in humanities, for example, might start from a more qualitative point of view.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[2] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep Variational Information Bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[4] Dan Amir and Ofra Amir. 2018. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1168–1176.

[5] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2018. Agent strategy summarization. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1203–1207.

[6] Yotam Amitai, Guy Avni, and Ofra Amir. 2023. ASQ-IT: Interactive Explanations for Reinforcement-Learning Agents. *arXiv preprint arXiv:2301.09941* (2023).

[7] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. 2019. Explaining reinforcement learning to mere mortals: An empirical study. *arXiv preprint arXiv:1903.09708* (2019).

[8] Guy Aridor, Rava Azeredo da Silveira, and Michael Woodford. 2023. Information-Constrained Coordination of Economic Behavior. (2023).

[9] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[10] Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing. 2021. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11396–11404.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[12] Yannick Brand and Axel Schulte. 2021. Workload-adaptive and task-specific support for cockpit crews: design and evaluation of an adaptive associate system. *Human-Intelligent Systems Integration* 3, 2 (2021), 187–199.

[13] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. 2019. Plan explanations as model reconciliation. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 258–266.

[14] Jessie YC Chen, Michael J Barnes, Anthony R Selkowitz, and Kimberly Stowers. 2016. Effects of agent transparency on human-autonomy teaming effectiveness. In *2016 IEEE international conference on Systems, man, and cybernetics (SMC)*. IEEE, 001838–001843.

[15] Jessie YC Chen, Michael J Barnes, Julia L Wright, Kimberly Stowers, and Shan G Lakhmani. 2017. Situation awareness-based agent transparency for human-autonomy teaming effectiveness. In *Micro-and nanotechnology sensors, systems, and applications IX*, Vol. 10194. SPIE, 362–367.

[16] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[17] Jessie Y Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. Situation awareness-based agent transparency. *US Army Research Laboratory* (2014), 1–29.

[18] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 81–87.

[19] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[20] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[21] Mica Endsley. 1995. Measurement of Situation Awareness in Dynamic Systems. *Human Factors* 37 (1995), 65–84.

[22] Mica R Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. IEEE, 789–795.

[23] Mica R Endsley. 2015. Situation awareness misconceptions and misunderstandings. *Journal of cognitive Engineering and Decision making* 9, 1 (2015), 4–32.

[24] Mica R Endsley. 2017. Direct measurement of situation awareness: Validity and use of SAGAT. In *Situational Awareness*. Routledge, 129–156.

[25] Mica R Endsley. 2023. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior* 140 (2023), 107574.

[26] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[27] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine* 40, 2 (2019), 44–58.

[28] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[29] Bradley Hayes and Julie A Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 303–312.

[30] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2022. It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 89–101.

[31] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[32] Mark K Ho. 2019. The value of abstraction. *Current opinion in behavioral sciences* 29 (2019).

[33] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[34] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[35] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. 2019. Enabling robots to communicate their objectives. *Autonomous Robots* 43 (2019), 309–326.

[36] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. 2021. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence* 301 (2021), 103571.

[37] Holly Huey, Xuanchen Lu, Caren M. Walker, and Judith E. Fan. 2023. Visual Explanations Prioritize Functional Properties at the Expense of Visual Fidelity. *Cognition* 236, C (2023), 105414.

[38] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. 2019. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*.

[39] Christopher Katins, Sebastian S Feger, and Thomas Kosch. 2023. Exploring Mixed Reality in General Aviation to Support Pilot Workload. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[40] Paul Kay, Brent Berlin, Luisa Maffi, William R Merrifield, and Richard Cook. 2009. *The world color survey*. CSLI Publications Stanford, CA.

[41] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[42] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).

[43] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *arXiv preprint arXiv:2301.09656* (2023).

[44] Michael S Lee, Henny Admoni, and Reid Simmons. 2022. Reasoning about Counterfactuals to Improve Human Inverse Reinforcement Learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9140–9147.

[45] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[46] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).

[47] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. 2019. Why these explanations? Selecting intelligibility types for explanation goals.. In *IUI Workshops*.

[48] Luca Longo. 2018. Experienced mental workload, perception of usability, their interaction and impact on task performance. *PloS one* 13, 8 (2018), e0199661.

[49] Luca Longo, Christopher D Wickens, Gabriella Hancock, and Peter A Hancock. 2022. Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology* 13 (2022), 883321.

[50] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[51] Eric J Michaud, Adam Gleave, and Stuart Russell. 2020. Understanding learned reward functions. *arXiv preprint arXiv:2012.05862* (2020).

[52] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[53] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[54] Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. 2021. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences* 118, 49 (2021), e2025993118.

[55] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems* 34 (2021), 610–623.

[56] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2008. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making* 2, 2 (2008), 140–160.

[57] Eugenio Piasini, Alexandre LS Filipowicz, Jonathan Levine, and Joshua I Gold. 2021. Embo: a Python package for empirical data analysis using the Information Bottleneck. *Journal of open research software* 9, 1 (2021).

[58] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[59] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=HPuSIXJaa9

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

[61] Stuart J Russell. 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.

[62] Sadiq Said, Malgorzata Gozdzik, Tadzio Raoul Roche, Julia Braun, Julian Rössler, Alexander Kaserer, Donat R Spahn, Christoph B Nöthiger, and David Werner Tscholl. 2020. Validation of the raw national aeronautics and space administration task load index (NASA-TLX) questionnaire to assess perceived workload in patient monitoring tasks: Pooled analysis study using mixed models. *Journal of medical Internet research* 22, 9 (2020), e19472.

[63] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 215–227.

[64] Lindsay Sanneman and Julie A Shah. 2022. An empirical study of reward explanations with human-robot interaction applications. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8956–8963.

[65] Lindsay Sanneman and Julie A Shah. 2022. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *International Journal of Human–Computer Interaction* 38, 18-20 (2022), 1772–1788.

[66] Lindsay Sanneman and Julie A Shah. 2023. Validating metrics for reward alignment in human-autonomy teaming. *Computers in Human Behavior* 146 (2023), 107809.

[67] Kristin E Schaefer, Edward R Straub, Jessie YC Chen, Joe Putney, and Arthur W Evans III. 2017. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research* 46 (2017), 26–39.

[68] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. 2021. Reward is enough. *Artificial Intelligence* 299 (2021), 103535.

[69] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. 2018. Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations.. In *IJCAI*. 4829–4836.

[70] DJ Strouse and David J Schwab. 2017. The deterministic information bottleneck. *Neural computation* 29, 6 (2017), 1611–1630.

[71] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[72] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. 2019. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 249–257.

[73] Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The Information Bottleneck Method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*.

[74] Heather L Tubbs-Cooley, Constance A Mara, Adam C Carle, and Ayse P Gurses. 2018. The NASA Task Load Index as a measure of overall workload among neonatal, paediatric and adult intensive care nurses. *Intensive and Critical Care Nursing* 46 (2018), 64–69.

[75] Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. 2022. Trading off Utility, Informativeness, and Complexity in Emergent Communication. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

[76] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence* (2020).

[77] Christopher D Wickens. 2008. Multiple resources and mental workload. *Human factors* 50, 3 (2008), 449–455.

[78] GN Wilkinson and CE Rogers. 1973. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 22, 3 (1973), 392–399.

[79] Julia L Wright, Jessie YC Chen, Michael J Barnes, and Peter A Hancock. 2016. Agent reasoning transparency's effect on operator workload. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 249–253.

[80] Julia L Wright, Stephanie A Quinn, Jessie YC Chen, and Michael J Barnes. 2014. Individual differences in human-agent teaming: An analysis of workload and situation awareness through eye movements. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 1410–1414.

[81] Ju-Chi Yu, Ting-Yun Chang, and Cheng-Ta Yang. 2014. Individual differences in working memory capacity and workload capacity. *Frontiers in psychology* 5 (2014), 1465.

[82] Mehrdad Zakershahrak, Ze Gong, Nikhillesh Sadassivam, and Yu Zhang. 2019. Online explanation generation for human-robot teaming. *arXiv preprint arXiv:1903.06418* (2019).

[83] Mehrdad Zakershahrak, Ze Gong, Nikhillesh Sadassivam, and Yu Zhang. 2020. Online explanation generation for planning tasks in human-robot teaming. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6304–6310.

[84] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences* 115, 31 (2018), 7937–7942.

[85] Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. 2021. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

[86] Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. 2019. Semantic categories of artifacts and animals reflect efficient coding. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

[87] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

[88] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

A GENERATING IB ABSTRACTIONS IMPLEMENTATION

We used the embo package to generate abstractions at different levels of complexity and reward distortion and employed these abstractions in human participant experiments to analyze the role of complexity and distortion in human understanding [57]. Here, we discuss how we generated our abstractions, and demonstrate that we induced meaningful variation in complexity and distortion.[1]

A.1 Grid Domain

In the grid domains, we numbered each of the 25 grid cells with a unique id, and used four reward functions during the IB process: Manhattan distance, random reward, x coordinate, and y coordinate. For the Manhattan distance reward, depicted in the main paper in Figure 1 a, reward was set to +1 at cell location $(1, 3)$ and decreased by 0.33 for each increase in Manhattan distance, to a minimum of -1.0. For the random reward, depicted in Figure 1 b, we selected a reward value uniformly at random in the range $[-1, 1]$. For exact values, we refer the reader to our code which, given the fixed random seed of 0, will exactly reproduce the random values in our experiments. Lastly, for the x and y coordinate rewards, we set the reward equal to the integer value of the x or y coordinate of each cell, ranging from 0 to 4, inclusive.

For each of the four training objects used during the IB process, we evaluated abstractions with the Manhattan and random grid reward functions. For example, we generated abstractions via the x coordinate reward function, which divided the grid into vertical regions, and evaluated distortion using such abstractions for the Manhattan and random grid rewards. Figure 2 a in the main paper shows how abstractions generated using different reward functions resulted in different distortion values for the same complexity.

A.2 Color Domain

In the color domain, we uniquely numbered each of the 122 colors in our experiments, and used three reward functions during the IB process: 1) predict the blue value in the color's RGB representation,

[1] Code used to generate the abstractions in our experiments is available at https://github.com/mycal-tucker/ib_xai.

2) predict the discontinuous reward (defined in the next paragraph) based on the blue value of the color, or 3) predict the red value of the color's RGB representation.

The continuous reward functions (predicting the blue or red value) were simply defined as the continuous value of the color, ranging from 0 to 1.0. We additionally used a discontinuous reward function, which divided the blue color range into eight bins of equal sizes (in $[0, 0.125)$, $[0.125, 0.25)$, etc.) with values of $[0.5, -0.5, 0.0, 0.75, 1.0, -1.0, 0.25, -0.75]$. We chose these values to intentionally cause similar continuous blue values (e.g., 0.124 and 0.126) to have dissimilar rewards (e.g., 0.5 and -0.5).

We include results of evaluating all such abstractions using the continuous blue reward function in Figure 6. Unsurprisingly, using the continuous blue reward to generate abstractions resulted in lower distortion for the same complexity compared with using the other two training objectives. Concretely, we observed that increasing the complexity of abstractions used for predicting a color's red value had virtually no effect on the distortion for predicting the color's blue value. Given the orthogonal nature of blue and red representations in RGB colors, this is unsurprising. Overall, we used these three training objects for generating abstractions to decouple changes in complexity and distortion, while exploring meaningful ranges for each value.

B COLOR DOMAIN RESULTS

In the main paper, we omitted some of the graphs from the color domain experiments for clarity; we include them here for completeness.

Figure 7 shows the key trends between Feature Rank and Distortion (a), Best Demonstration and Distortion (b), and Workload and Complexity (c) using the discontinuous reward function. In the main paper, we noted that the FR-distortion trend was significant but the Workload-Complexity trend was not, while the Best Demonstration results were never significant in any color domain. The plots in Figure 7 visually corroborate these statistical tests. The Best Demonstration trend line is nearly flat, including some participants predicting the exact right demonstration at a distortion of over 0.4, which indicates high performance using uninformative abstractions. At the same time, the trend between complexity and workload is positive, as we hypothesized.

C ABSTRACTION VISUALIZATIONS

Throughout our experiments, we used abstractions at different complexity and distortion levels and typically reported the complexity and distortion metrics. Here, to provide intuition about the types of abstractions used, we provide visualizations for both domains, at various complexity and distortion values.

Figures 8 and 9 include visualizations of abstractions for the Manhattan and random grid domains, respectively. Within each figure, we selected three checkpoints at low, medium, and high complexity (corresponding to different columns). Note how the low-complexity checkpoints used fewer abstractions than higher-complexity checkpoints.

The different rows in each figure reflect abstractions generated using different training objectives in the IB process. Recall from Section 4.2 that we used different training objectives when generating abstractions to induce different distortions for the same complexity. Abstractions based on different reward functions are included as different rows in Figures 8 and 9. For example, the top row of Figure 8 shows abstractions generated using the Manhattan reward function; the second row, however, depicts abstractions generated using the x coordinate of each location. Even as the x-based abstractions increased in complexity, until each x value was represented separately, such abstractions remained poor for predicting the actual Manhattan reward in the grid. Similar patterns were true in the random grid as well (Figure 9). Overall, the visualizations of abstractions in these grid domains confirm intuitions that 1) increasing complexity led to finer-grained abstractions and 2) abstractions generated using sub-optimal training objectives led to poor reward prediction, and therefore high distortion.

We note that our visualizations of the abstractions are merely one possible way of representing the underlying abstraction. In general, we consider an abstraction as a lossy, discrete representation that captures one or more inputs. In Figures 8, therefore, we show how different grid locations (inputs) are grouped into different abstractions that produce different rewards (represented as colors in the visualizations). Even as our definition of abstractions remains fixed, one may consider alternate depictions of abstractions, such as sampling only a subset of inputs that map to a given abstraction.

Figure 10 includes visualizations of abstractions from the color domain much like the previous visualizations of grid-based abstractions. Abstractions were evaluated according to the continuous blue reward function; the heatmap used for visualization shows the average blue value of each abstraction. As before, we selected three checkpoints at low, medium, and high complexity (corresponding to different columns).

The different rows in Figure 10 reflect abstractions generated using different training objectives. In the top row, we used the continuous blue reward function – the same function used for evaluation. In the second row, we used the discontinuous blue reward function, and in the third row we used a color's red value to generate abstractions.

As in the grid domains, we found that 1) increasing complexity led to more fine-grained abstractions and 2) using different training objectives in the IB process led to sub-optimal abstractions that resulted in high distortion. For example, using the discontinuous blue reward function (second row) at high complexity (Figure 10 f), we observed that two abstractions have nearly identical average rewards of 0.17 and 0.18. Thus, the additional complexity incurred by having more abstractions comes with barely any decrease in distortion. This suboptimality is even more pronounced for abstractions generated with the continuous red reward function (bottom row). Using just two abstractions at low complexity (Figure 10 g), the color space is partitioned into two groups that are meaningful for predicting the redness of a color, but with almost no difference in blue value (mean values of 0.43 and 0.47).

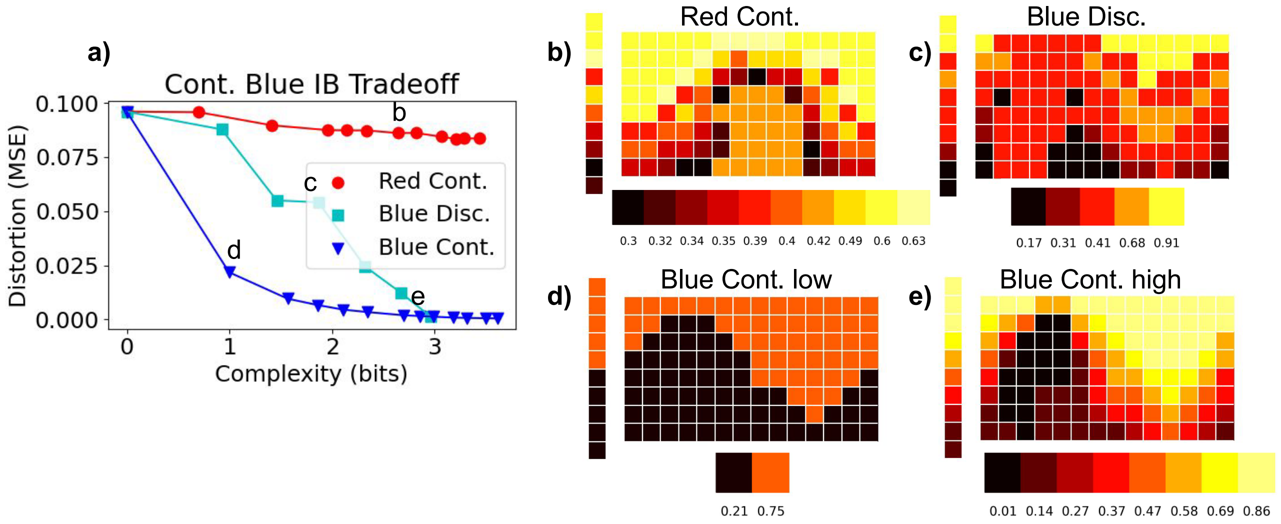
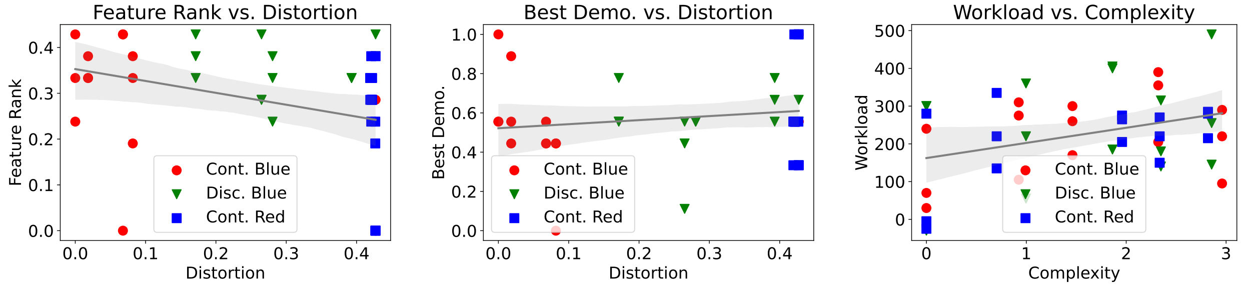


Figure 6: Complexity-distortion curves (a)) and corresponding abstractions (b-e) for the continuous reward function in the color domain. Using different rewards to generate abstractions (different curves) led to varying distortions for the same complexity. Using the continuous reward function led to optimal distortion-complexity tradeoffs, and varying complexity increased the number of abstractions (d-e).



(a) Feature Rank vs. Distortion

(b) Best Demonstration vs. Distortion

(c) Workload vs. Complexity

Figure 7: Color domain results using the discontinuous blue reward function. Trends were weaker than those observed with the continuous reward, although there was still a significant negative correlation between feature rank and distortion (a).

D SURVEY QUESTIONS

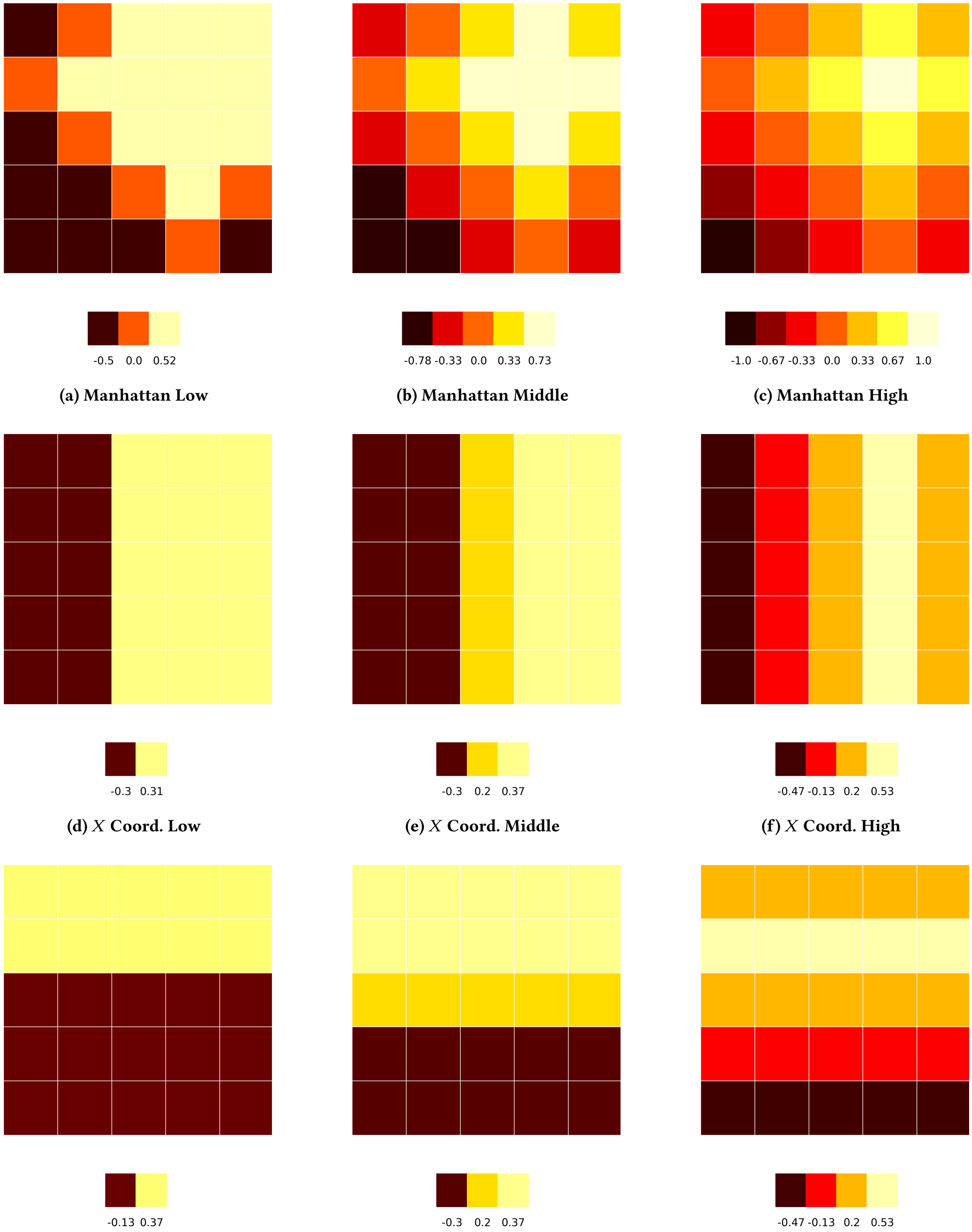
Here, we include examples of each question asked of participants in the grid navigation survey and the color survey. Each participant answered two of each question within the particular survey they were responding to: one for a continuous reward function and one for a discontinuous reward function.

D.1 Grid Navigation Domain

Figures 11 and 12 depict the *feature rank (FR)* and the *best demonstration (BD)* questions as they were presented to participants in the grid navigation domain. Here, the abstract explanation shown in each question (via a heat map indicating the values of the grid squares) represents the ground truth reward function. In this case, grid squares in the upper left corner have values of +1.0, those in the upper right have values of +0.5, those in the middle row have values of 0.0, those in the lower left have values of -0.5, and those in the lower right have values of -1.0.

D.2 Color Domain

Figures 13 and 14 depict the *feature rank (FR)* and the *best demonstration (BD)* questions as they were presented to participants in the color domain. The abstract explanation shown in each question (via a heat map indicating the values of each color in the color grid) represents the ground truth reward function. Colors in the top part

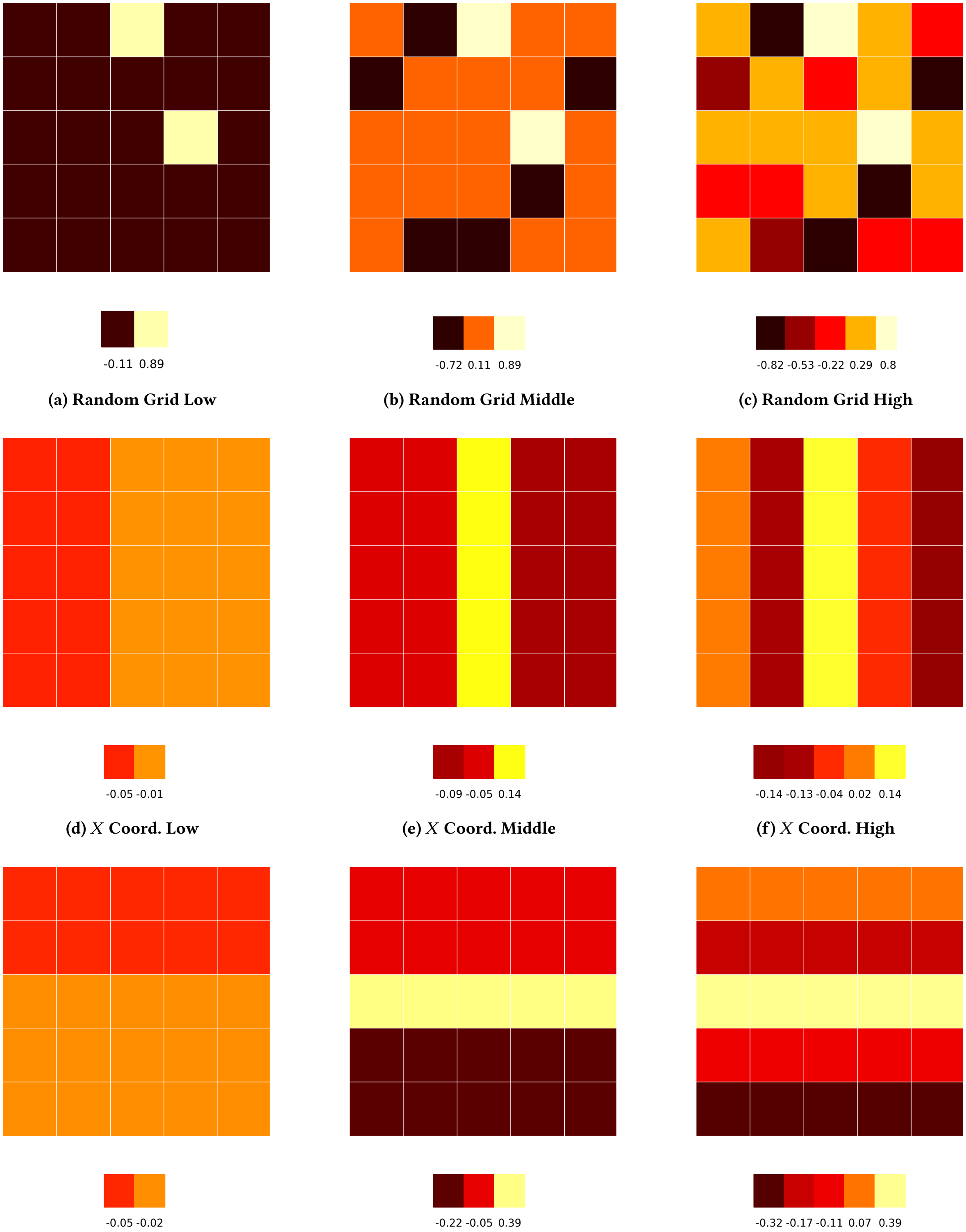


(a) Manhattan Low **(b) Manhattan Middle** **(c) Manhattan High**

(d) X Coord. Low **(e) X Coord. Middle** **(f) X Coord. High**

(g) Y Coord. Low **(h) Y Coord.Middle** **(i) Y Coord. High**

Figure 8: Manhattan grid abstractions for various training objectives (different rows) and complexity levels (different columns). Increasing complexity led to more and finer-grained abstractions. Some abstractions led to low distortion (top row), whereas others removed important information, leading to high distortion (bottom two rows).



(a) Random Grid Low

(b) Random Grid Middle

(c) Random Grid High

(d) X Coord. Low

(e) X Coord. Middle

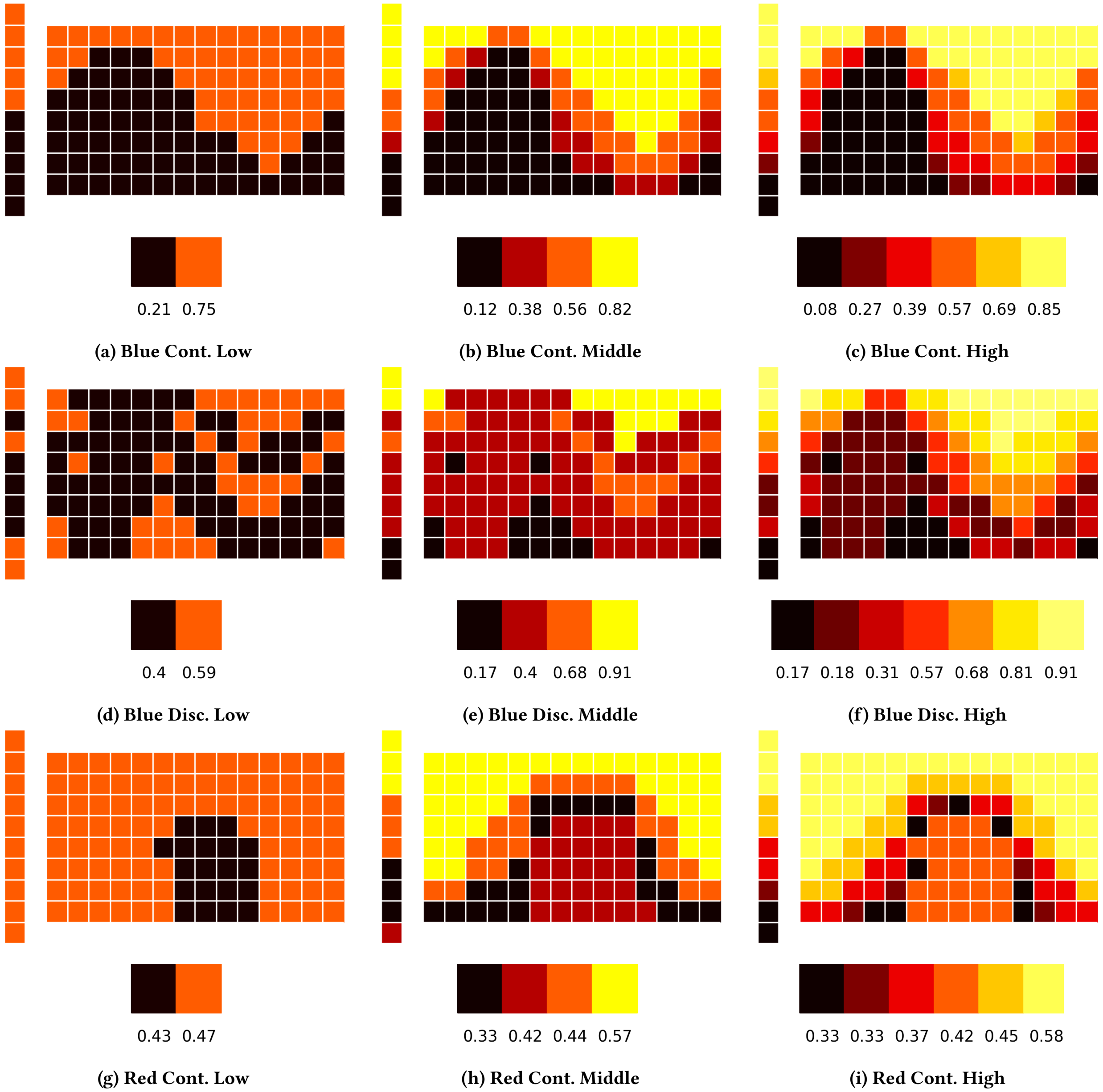
(f) X Coord. High

(g) Y Coord. Low

(h) Y Coord. Middle

(i) Y Coord. High

Figure 9: Random grid abstractions for various training objectives (different rows) and complexity levels (different columns). As in the Manhattan grid abstractions, more complex abstractions captured finer-grained information, but the distortion for a given complexity level depended upon the training objective.



(a) Blue Cont. Low **(b) Blue Cont. Middle** **(c) Blue Cont. High**

(d) Blue Disc. Low **(e) Blue Disc. Middle** **(f) Blue Disc. High**

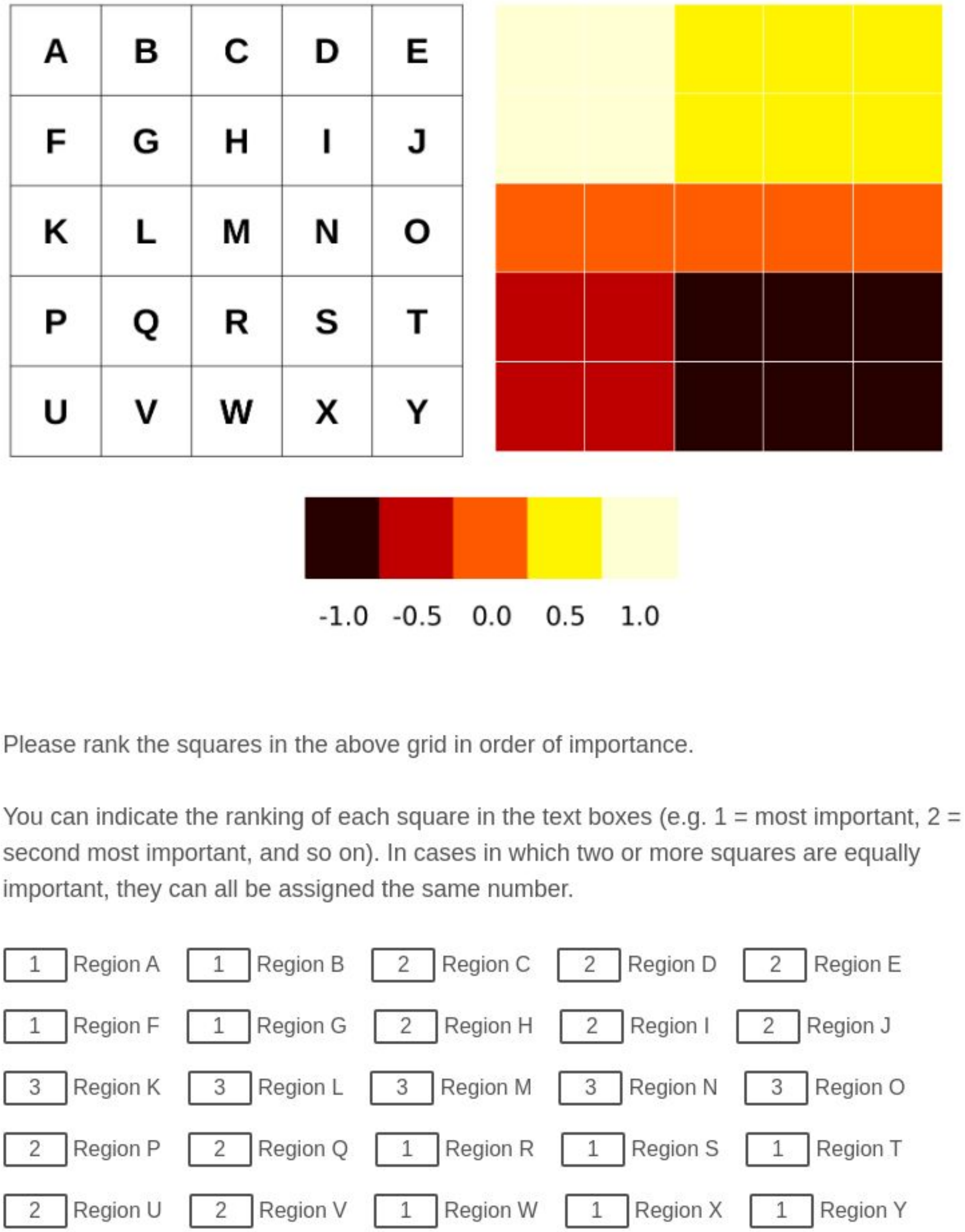
(g) Red Cont. Low **(h) Red Cont. Middle** **(i) Red Cont. High**

Figure 10: Color abstractions evaluated on the continuous blue reward function. Increasing complexity (left to right) increased the number of abstractions. The distortion associated with such abstractions varied greatly, however, depending upon the training objective used when generating abstractions. Using the continuous blue reward to generate abstraction (top row) led to evenly spaced abstractions with little distortion. Different training objectives (discontinuous blue in the second row, continuous red in the third) led to suboptimal abstractions with higher distortion.

of the grid have values of +0.5, those below that have values of +1.0, those below that have values of -1.0, and those at the bottom have values of -0.5.

D.3 Workload Questions

The workload questions that we applied in our surveys were drawn from the NASA TLX survey [28]. The set of questions as they were displayed in the grid navigation domain are depicted in Figure 15. The questions shown to participants in the color domain were nearly identical with minor wording changes based on the differences between the two domains.



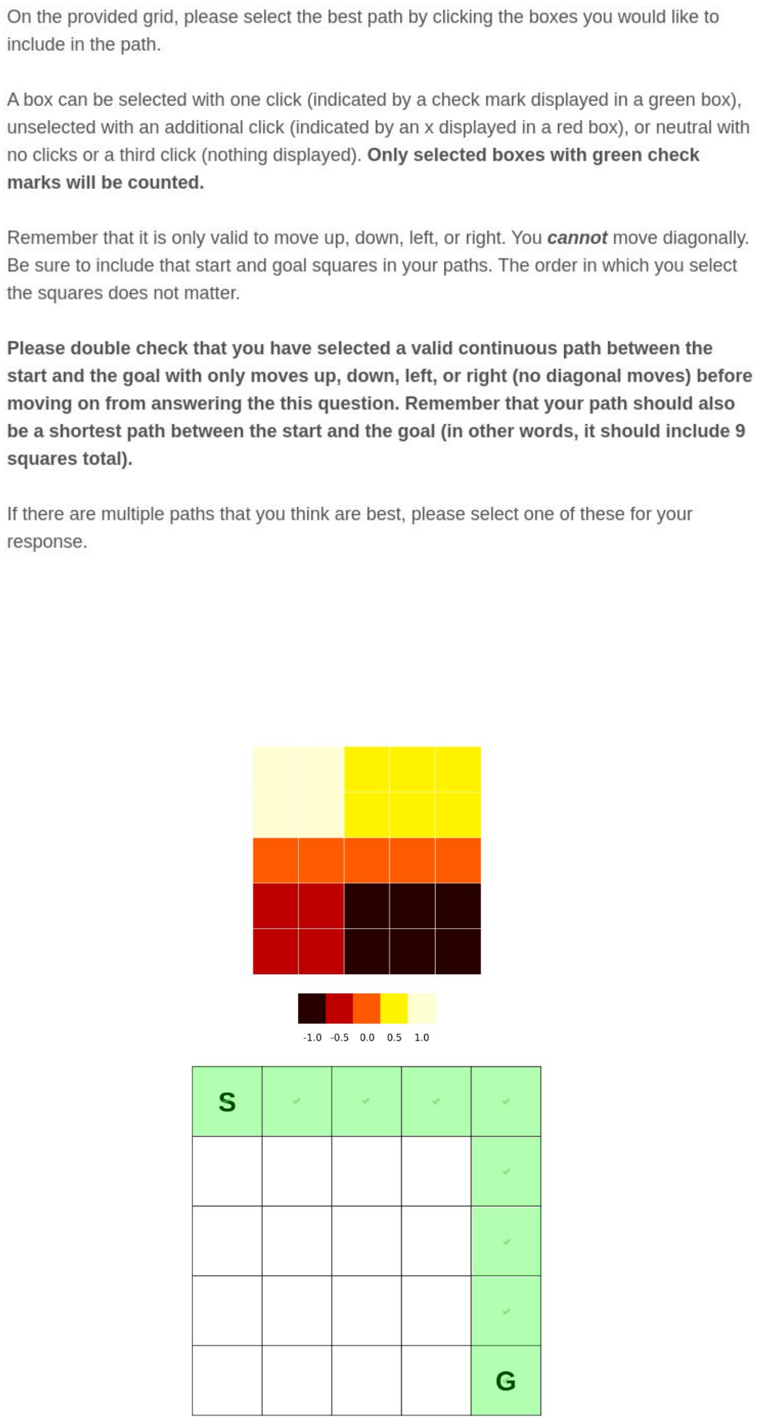
A	B	C	D	E
F	G	H	I	J
K	L	M	N	O
P	Q	R	S	T
U	V	W	X	Y

Please rank the squares in the above grid in order of importance.

You can indicate the ranking of each square in the text boxes (e.g. 1 = most important, 2 = second most important, and so on). In cases in which two or more squares are equally important, they can all be assigned the same number.

1 Region A	1 Region B	2 Region C	2 Region D	2 Region E
1 Region F	1 Region G	2 Region H	2 Region I	2 Region J
3 Region K	3 Region L	3 Region M	3 Region N	3 Region O
2 Region P	2 Region Q	1 Region R	1 Region S	1 Region T
2 Region U	2 Region V	1 Region W	1 Region X	1 Region Y

Figure 11: Example of the *feature rank* question in the grid navigation domain. The top left grid designates regions of the grid which are ranked by participants when answering the question. The top right grid depicts the abstract reward regions. The reward values associated with each reward region are depicted in the five numbered swatches below the grids. At the bottom, sample responses for the *feature rank* question are provided based on the given abstract grid.



On the provided grid, please select the best path by clicking the boxes you would like to include in the path.

A box can be selected with one click (indicated by a check mark displayed in a green box), unselected with an additional click (indicated by an x displayed in a red box), or neutral with no clicks or a third click (nothing displayed). **Only selected boxes with green check marks will be counted.**

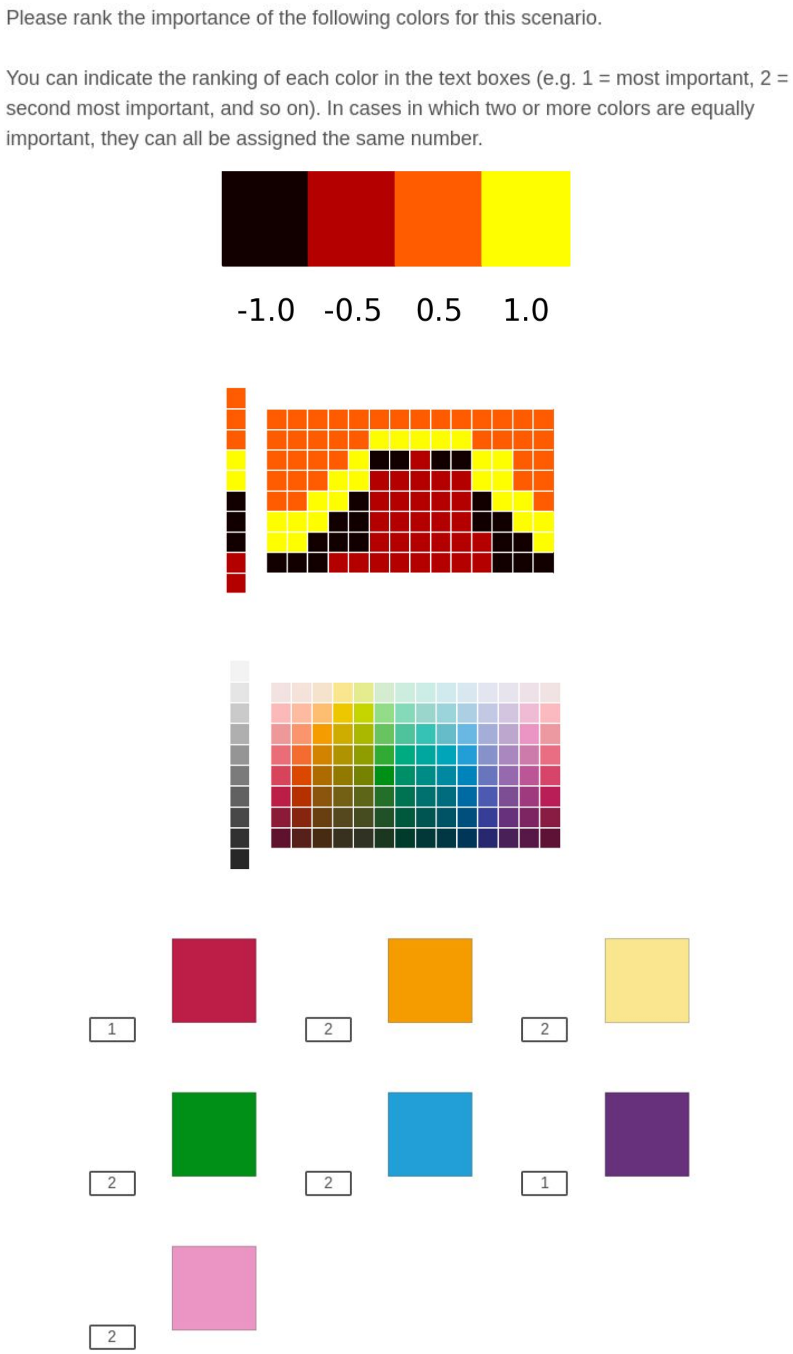
Remember that it is only valid to move up, down, left, or right. You ***cannot*** move diagonally. Be sure to include that start and goal squares in your paths. The order in which you select the squares does not matter.

Please double check that you have selected a valid continuous path between the start and the goal with only moves up, down, left, or right (no diagonal moves) before moving on from answering the this question. Remember that your path should also be a shortest path between the start and the goal (in other words, it should include 9 squares total).

If there are multiple paths that you think are best, please select one of these for your response.

S	✓	✓	✓	✓
				✓
				✓
				✓
				G

Figure 12: Example of the *best demonstration* question in the grid navigation domain. The same explanation (grid with the abstract reward regions) is shown to participants as for the *feature rank* question. For this question, participants must select the reward-maximizing shortest path through the grid at the bottom of the figure (depicted through the selected green boxes with check marks).



Please rank the importance of the following colors for this scenario.

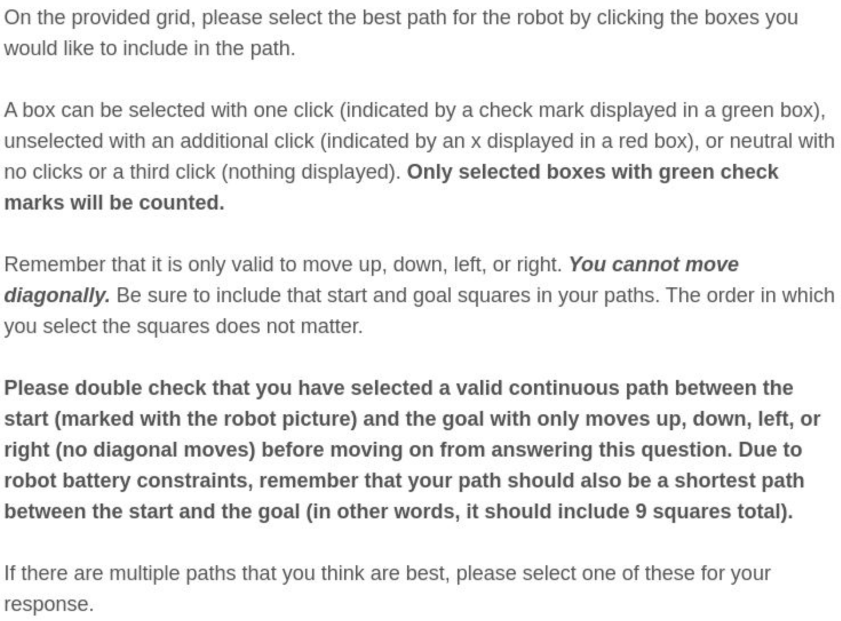
You can indicate the ranking of each color in the text boxes (e.g. 1 = most important, 2 = second most important, and so on). In cases in which two or more colors are equally important, they can all be assigned the same number.

1 2 2

2 2 1

2

Figure 13: Example of the *feature rank* question in the color domain. The heat map and the associated numbered color swatches at the top indicate the abstract regions of colors from the color grid below and their corresponding reward values. At the bottom, sample responses for the *feature rank* question (a set of ranked colors from the color grid) are provided based on the given abstraction of the color grid.



On the provided grid, please select the best path for the robot by clicking the boxes you would like to include in the path.

A box can be selected with one click (indicated by a check mark displayed in a green box), unselected with an additional click (indicated by an x displayed in a red box), or neutral with no clicks or a third click (nothing displayed). **Only selected boxes with green check marks will be counted.**

Remember that it is only valid to move up, down, left, or right. ***You cannot move diagonally.*** Be sure to include that start and goal squares in your paths. The order in which you select the squares does not matter.

Please double check that you have selected a valid continuous path between the start (marked with the robot picture) and the goal with only moves up, down, left, or right (no diagonal moves) before moving on from answering this question. Due to robot battery constraints, remember that your path should also be a shortest path between the start and the goal (in other words, it should include 9 squares total).

If there are multiple paths that you think are best, please select one of these for your response.

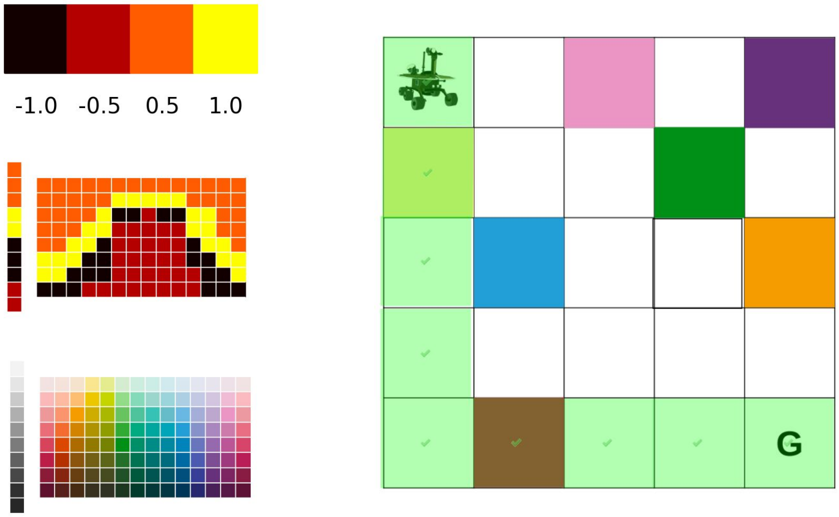
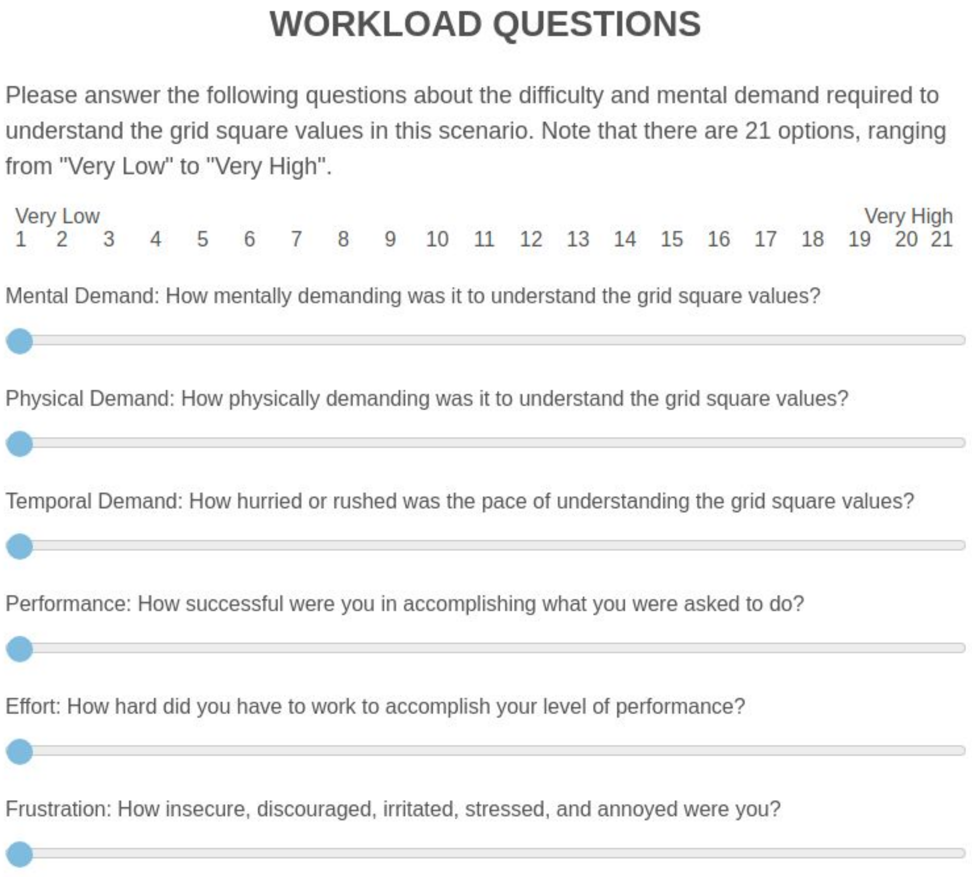


Figure 14: Example of the *best demonstration* question in the color domain. The same explanation (color grid with the abstract reward regions) is shown to participants as for the *feature rank* question. Here, participants must select the path through the grid at the bottom of the figure which maximizes the value of the collected samples, which are indicated by different colors from the original color grid. The selected path is, again, marked by green boxes with check marks. Note that as in the grid navigation domain, participants were asked to identify a shortest path through the grid.



WORKLOAD QUESTIONS

Please answer the following questions about the difficulty and mental demand required to understand the grid square values in this scenario. Note that there are 21 options, ranging from "Very Low" to "Very High".

Very Low Very High
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Mental Demand: How mentally demanding was it to understand the grid square values?

Physical Demand: How physically demanding was it to understand the grid square values?

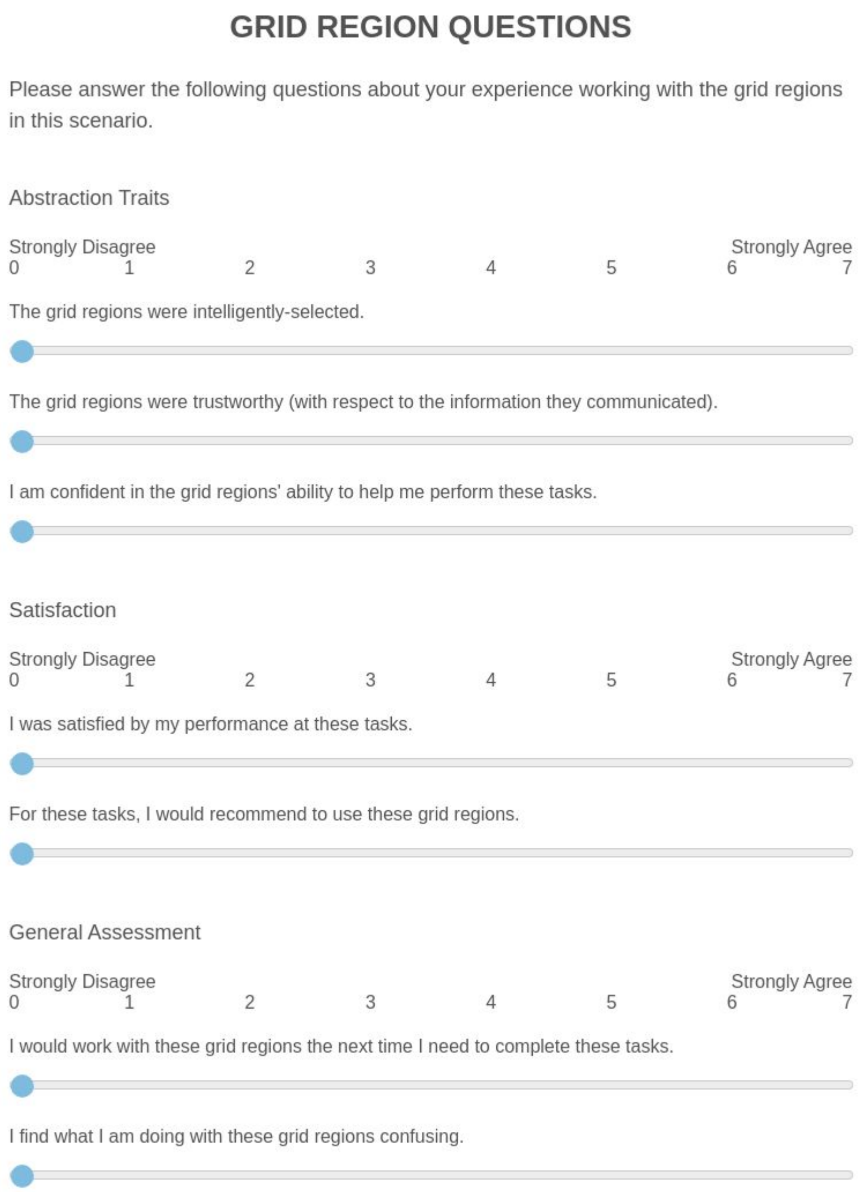
Temporal Demand: How hurried or rushed was the pace of understanding the grid square values?

Performance: How successful were you in accomplishing what you were asked to do?

Effort: How hard did you have to work to accomplish your level of performance?

Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

Figure 15: Questions from the NASA TLX survey [28]. The questions above are depicted for the grid navigation domain.



GRID REGION QUESTIONS

Please answer the following questions about your experience working with the grid regions in this scenario.

Abstraction Traits

Strongly Disagree Strongly Agree
0 1 2 3 4 5 6 7

The grid regions were intelligently-selected.

The grid regions were trustworthy (with respect to the information they communicated).

I am confident in the grid regions' ability to help me perform these tasks.

Satisfaction

Strongly Disagree Strongly Agree
0 1 2 3 4 5 6 7

I was satisfied by my performance at these tasks.

For these tasks, I would recommend to use these grid regions.

General Assessment

Strongly Disagree Strongly Agree
0 1 2 3 4 5 6 7

I would work with these grid regions the next time I need to complete these tasks.

I find what I am doing with these grid regions confusing.

Figure 16: Subjective assessment questions adapted from Hoffman [33]. The questions above are depicted for the grid navigation domain, and were largely similar in the color domain.

D.4 Subjective Assessment Questions

The subjective assessment questions that we applied in our surveys were adapted from the scale for team fluency proposed by Hoffman [33]. The set of questions as they were displayed in the grid navigation domain are depicted in Figure 16. The questions shown to participants in the color domain were, again, nearly identical with minor wording changes based on the differences between the two domains.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009