Abstract—This paper investigates the impact of framing and time pressure on human judgment performance in a complex multiattribute judgment task. We focus on the decision process of human participants who must choose between pairwise alternatives in a resource-allocation task. We used the Analytic Hierarchy Process (AHP) to calculate the relative weights of the four alternatives (i.e., C_1, C_2, C_3, and C_4) and the judgment consistency. Using the AHP, we examined two sets of hypotheses that address the impact of task conditions on the weight prioritization of choice alternatives and the internal consistency of the judgment behavior under varying task conditions. The experiment simulated the allocation of robotic assets across the battlefield to collect data about an enemy. Participants had to make a judgment about which asset to allocate to a new area by taking into account three criteria related to the likelihood of success. We manipulated the information frame and the nature of the task. We found that, in general, participants gave significantly different weights to the same alternatives under different frames and task conditions. Specifically, in terms of ln-transformed priority weights, participants gave significantly lower weights to C_2, lower weights to C_3 in Task #3, and the judgment behavior under varying task conditions. The experiment simulated the allocation of robotic assets across the battlefield to collect data about an enemy. Participants had to make a judgment about which asset to allocate to a new area by taking into account three criteria related to the likelihood of success. We manipulated the information frame and the nature of the task. We found that, in general, participants gave significantly different weights to the same alternatives under different frames and task conditions. Specifically, in terms of ln-transformed priority weights, participants gave significantly lower weights to C_2 and C_4, and higher weight to C_3 in Task #3, participants gave significantly higher weight to C_2 in Task #1, lower weights to C_1 and C_4, higher weight to C_3 in Task #2, and lower weight to C_1 in Task #3. Furthermore, we found that the internal consistency of the decision behavior was worse, first, in the loss frame than the gain frame and, second, under time pressure. Our methodology complements utility-theoretic frameworks by assessing judgment consistency without requiring the use of task-performance outcomes. This work is a step toward establishing a coherence criterion to investigate judgment under naturalistic conditions. The results will be useful for the design of multiattribute interfaces and decision aiding tools for real-time judgments in time-pressured task environments.

Index Terms—Analytic Hierarchy Process (AHP), framing effect, judgment consistency, multiattribute task.

I. INTRODUCTION

This paper evaluates the judgment effectiveness of human participants in a dynamic resource-allocation task in a command-and-control context. Specifically, we focus on the effect of information framing and task complexity on judgment performance. Framing refers to the manner in which a choice problem is presented, and its effects on the decision processes have been demonstrated in judgment and decision-making literature [1]. Framing effects have been found in loss/gain and sunk-cost problems [2], [3], in a multiattribute dynamic decision-making environment [4], and across distinct task domains [5]. More recently, its effects have also been investigated in medical decision-making processes [6], in financial decision-making processes [7]–[9], and in a risk-elicitation process [10].

Framing has been characterized in diverse manners in the decision literature. Most decision researchers describe framing effects in terms of deviation from an expected value that is based on a decision maker’s utility function. For example, Kahneman and Tversky [11] demonstrated that framing outcomes as losses versus gains can cause reversals in the selected choice through using a definition of risk as the deviation from the expected utility. Whereas research efforts that rely upon utility theory are mathematically rigorous, the derived results may not be representative of human capabilities [12]. In the light of decision-making research which suggests that humans derive solutions that are satisfactory rather than optimal [13], [14] and that humans utilize simple heuristics rather than elaborate computation methods [15], [16], utility-theoretic representations of human judgment have become untenable.

In addition to objections about assumptions on human capabilities, there exists the difficulty of establishing a coherence criterion for truth [17] in representative environments, i.e., it is difficult to assess the internal consistency of the observed behavior with axioms prescribed by some normative model when no normative model has been explicitly adopted by the research community.

Therefore, we propose a testable coherence criterion to address both information framing and task complexity. We designed our experimental tasks based on Nygren’s work [4] in which participants performed three multiattribute dynamic decision-making tasks (i.e., data-tracking, system-monitoring, and resource-management tasks) under different framing conditions of the payoff scheme (e.g., accumulating performance points under positive framing and losing performance points under negative framing), and their task performances, such
as type of error, reaction time, and number and pattern of responses, were measured in terms of accuracy and latency. While Nygren [4] provides an analysis of task performance measures subject to framing effects, we investigate judgment performance subject to information framing and task complexity manipulations. We consider whether participants make consistent judgments within a specific frame, across loss and gain frames, and across tasks of different complexity. We will address consistency within a set of judgments as well as the persistence of weight prioritization under different task manipulations.

II. PROBLEM STATEMENT

Military robots have revolutionized the battlefield as sensors and weapons [18], [19], and the assessment of human allocation strategies for limited assets is an area of active research. Fig. 1 shows a graphical representation of the problem. A regional commander specifies Named Areas of Interest (NAIs) that need to be monitored for enemy activity for a particular time interval. Given that a new NAI has just been specified, the operator must rearrange the existing robotic assets, shown as wedges, to create a new asset configuration that has sufficient coverage on all the NAIs. In this problem, the coverage sufficiency of an asset is defined along three attributes. First, the timeliness attribute addresses the probability that the asset will reach the new NAI within the defined time interval. Second, the longevity attribute concerns the probability that the asset will be able to collect data for the duration of the time interval. Third, the expected coverage attribute deals with the probability that the overall area is covered by all the assets.

The problem is hierarchically shown in Fig. 2. The primary objective of selecting a better asset configuration can be assessed using multiple attributes (i.e., timeliness, longevity, and expected coverage), which can, in turn, be decomposed into specific configurations (spatial arrangements of the asset, with each configuration having its own attribute values).

III. USING THE AHP TO MEASURE JUDGMENT

In utility theory, a key problem is the elicitation of the utility or worth of an attribute along some meaningful measurement scale. The problem is compounded in decisions involving multiple attributes. Therefore, a workable definition of consistency along the lines of the normative axioms of decision making is not tenable. To circumvent the difficulties associated with utility theory, we used the Analytic Hierarchy Process (AHP).

The AHP is a general theory of measurement used to derive ratio scales from human judgments through multiple paired comparisons in a hierarchically structured task [20]. The method of paired comparisons is not a new psychophysical technique (for an introductory treatment, see [21]) and is a foundational element for modern psychophysics [22]. The AHP has been used to assist decision making in applications ranging from bridge-design selection to product-pricing-strategy choice [23]. It has also been used in human-factor applications such as to rank order computer interfaces [24], to ascertain appropriate knowledge-elicitation methods [25], to select attributes for designing virtual environment systems [26], to isolate Gestalt grouping principles to organize home pages [27], and to analyze the decision process itself in multiattribute decisions [28]. While a brief introduction to the AHP is presented in this paper, a more comprehensive tutorial on using the AHP is provided by Mitta [24].

The first step in the AHP is to model the problem as a hierarchy, such as that shown in Fig. 2, containing the decision objective, the alternatives for reaching it, and the attributes for evaluating the alternatives. The next step is to establish priority weights among the elements (i.e., alternatives and attributes in Fig. 2) of the hierarchy by making a series of judgments based on pairwise comparisons of the elements. More specifically, participants make \( n(n - 1)/2 \) pairwise comparisons for \( n \) elements using a nine-point scale ranging from absolute dominance to equality. The results of the pairwise comparisons are then arranged in a reciprocal matrix

\[
A = \begin{bmatrix}
1 & w_1/w_2 & \cdots & w_1/w_n \\
w_2/w_1 & 1 & \cdots & w_2/w_n \\
\vdots & \vdots & \ddots & \vdots \\
w_n/w_1 & w_n/w_2 & \cdots & 1
\end{bmatrix}
\]

(1)

where \( w_j \) (\( j = 1, 2, \ldots, n \)) is the weight value of element \( j \).

Given a reciprocal matrix, the next step is to compute a weight vector \( w \), which represents the relative weights of the elements of the hierarchy. Saaty [20] showed that \( w \) must satisfy

\[
Aw = \lambda_{\text{max}} w
\]

(2)

where \( \lambda_{\text{max}} \) is the largest eigenvalue of \( A \). Mathematically, \( w \) can be obtained by normalizing the principle eigenvector of \( A \).
For a multilevel hierarchy, we can establish a hierarchical vector of weights \( w_i^h \) where \( i \) is an element of the hierarchy in a level \( h - 1 \). Each vector of weights can then be aggregated into a performance matrix for each level of the hierarchy \( W^h \). Finally, a performance matrix containing priority weights of the lowest level of the hierarchy with respect to the overall objective \( U \) can be calculated as

\[
U = W^v, W^{v-1}, \ldots, W^2
\]

(3)

where \( v \) is the number of levels in the hierarchy. Details of the derivation can be found in [20] and [28].

In addition to calculating priority weights, the AHP is also used to assess judgment consistency. Since small changes in a reciprocal matrix element imply a small change in \( \lambda_{\text{max}} \), the deviation of the latter from \( n \) can be taken as a measure of consistency. Saaty [20] uses this deviation as a measure he calls the consistency index, where

\[
CI = \frac{\lambda_{\text{max}} - n}{n-1}
\]

(4)

In other words, a set of choices is most consistent if \( CI = 0 \). The \( CI \) can then be compared with a random index (\( RI \)), which is the \( CI \) value of a randomly generated comparison matrix, to determine if any inconsistency found is acceptable. This is called the consistency ratio (\( CR \)). \( CR \) is then calculated as follows, and the \( RI \) values for a different number of \( n \) are shown in Table I:

\[
CR = \frac{CI}{RI}
\]

(5)

Once again, for a multilevel hierarchy, an aggregate \( CR \) can be obtained by considering the \( CI \) and \( RI \) of each level. For instance, the \( CI \) of a three-level hierarchy \( M \), like ours, can be calculated as

\[
M = CI^2 + W^2
\]

(6)

where \( CI^2 \) is the \( CI \) value of the second level and \( CI_i^3 \) is the \( CI \) value for the \( i \)th element of the third level for \( m \) elements. Similarly, the \( RI \) for a three-level hierarchy \( M \) can be calculated as

\[
\bar{M} = RI^2 + W^2
\]

(7)

where \( RI^2 \) is the \( RI \) value for the number of elements in the second level and \( RI_i^3 \) is the \( RI \) value for the number of elements in the third level. Finally, the consistency ratio of the hierarchy (\( CRH \)) can be calculated as

\[
CRH = \frac{M}{\bar{M}}
\]

(8)

As with \( CR \), consistency in judgment increases as \( CRH \) decreases.

For readers who are not familiar with the AHP procedure, a thorough example of demonstrating the use of (1)–(8) for our problem is provided in the Appendix.

We acknowledge that the AHP use has been controversial. Specifically, Dyer [29], [30] claimed that the AHP is flawed as a procedure because of the phenomenon of rank reversal in which an added alternative may cause the rank of two existing alternatives to be reversed. Harker and Vargas defended the AHP on the grounds that rank reversal is not a flaw because it is consistent with the nature of relative measurement [31] and that the strength of the attribute should be preserved, not the choice alternatives [32]. Nevertheless, Harker and Vargas [33] do stipulate that alternatives at level \( h \) of the hierarchy must be independent from attributes at level \( h - 1 \). This stipulation holds for our experiments. Dyer [30] and Harker and Vargas [33] provide additional viewpoints.

IV. RESEARCH HYPOTHESES

In this paper, we focus on the decision process of human judges who must choose between pairwise alternatives in a resource-allocation task. We regard the persistence of weight prioritization as the similarity of priority weights of choice alternatives over different experimental treatments. Moreover, we regard judgment consistency as the \( CRH \) value of a judge. Using the AHP, we examined the following research hypotheses.

1) Weight prioritization remains the same under gain or loss framing (H1).
2) Weight prioritization remains the same under different task types (H2).
3) Judgment consistency remains the same under gain or loss framing (H3).
4) Judgment consistency remains the same under different task types (H4).

The first two hypotheses deal with the impacts of framing and task conditions on the weight prioritization of alternatives. The second two hypotheses address the internal consistency of the decision behavior under varying task conditions.

V. EXPERIMENTAL DESIGN AND PROCEDURE

The experiment utilized a two-factorial within-subjects design. One factor was framing and consisted of a gain frame and a loss frame. The other factor was task type and included a static task (Task #1), a human-in-the-loop simulation task (Task #2), and a human-in-the-loop simulation task under time constraint (Task #3). Because of task differences and the amount of time
pressure applied, our unit of analysis is the outcome of a participant under a specific framing condition. Each outcome takes two forms: 1) priority weights of four alternatives and 2) a CRH value.

Forty-eight Pennsylvania State University undergraduate students (19 women and 29 men) participated in the experiments. The students ranged in age from 19 to 24 years ($\mu = 21.6$ years). The experiment took approximately 2 h to complete, and participants were compensated with $20.

Each of the 48 participants performed the required tasks under all six conditions of the two factors (i.e., $2 \times 3 = 6$). To avoid any learning or carryover effects, we counterbalanced the experimental design such that half of the participants were assigned to the gain frame first and then to the loss frame and that the other half of the participants were assigned to the loss frame first and then to the gain frame. To further reduce learning effects, the order of tasks presented to the participants under each frame was randomized.

For each task presentation, the participant made a choice between two configurations of robotic asset allocation. Participants were instructed that robotic assets have already been placed across the battlefield to cover a specific NAI (see Fig. 1). However, a new NAI has been identified, and because of time constraints, there is only time enough to move one asset to the new NAI. The system recommends the two best assets to be moved to the new NAI, and the human decides which asset to move. In deciding between configurations, participants needed to consider coverage sufficiency per attribute listed in Table II. The first three and the last three attributes in the table were used to represent the coverage sufficiency of an asset configuration in the gain and loss frames, respectively. In essence, there exist probabilistic relations among the attributes, such as $U = 1 - T$, $D = 1 - L$, and $EV = 1 - EC$. For instance, if the Timeliness of an asset was 0.67 in the gain frame, then the Untimeliness of the same asset was $1 - 0.67 = 0.37$ in the loss frame. These relations were used to frame the same asset configurations as gains or losses in the actual experiment.

To elicit the participants’ preference, we established two possible values per attribute, namely, average and high in the gain frame (or average and low in the loss frame), and generated only four combinations per frame, as shown in Table III. While the number of possible combinations of the values of the three attributes per frame is eight (i.e., $2^3 = 8$), we used only four of them because of the following: 1) the other possible four combinations, which are not included in Table III, are dominant over the generated four combinations, and 2) if they were included, it would drastically increase the number of pairwise comparisons for the generated configurations.

Based on the levels of the three attributes, we termed the four combinations as $C_1$–$C_4$ in the gain frame (and as $C_7$–$C_8$ in the loss frame). In essence, considering the probabilistic relations aforementioned, $C_1$–$C_4$ and $C_7$–$C_8$ are identical, but are framed as gains and losses, respectively.

Since the normal distribution has been widely used to represent the lifetime of items or systems (e.g., Longevity in our problem), which exhibit failure due to wear or malfunction, and also the variations in value (e.g., Timeliness and Expected Coverage in our problem) of electronic components or mechanical system, we utilized the normal distribution as the probability distribution of the three attributes and randomly generated the average values from a normal distribution $N(\mu = 0.50, \sigma = 0.0567)$, the high values in the gain frame from $N(\mu = 0.83, \sigma = 0.0567)$, and the low values in the loss frame from $N(\mu = 0.17, \sigma = 0.0567)$.

### A. Task #1

In Task #1, participants made a preference judgment on a nine-point scale between two asset configurations such as the ones shown in Fig. 3. The display showed a gain-framing condition and consisted of a map of the battlefield (left upper), an intelligence window that provided some detailed information about each scenario (left bottom), a decision aid that showed the two candidate configurations and the values of three attributes associated with the two configurations (right upper), and a question to elicit the participants’ preference between the two asset configurations (lower right). This display was static, and participants had no time limit to complete each question.

Note that the question posed in Fig. 3 looks like selecting one of the two sensor assets (ARV-1 and MULE-1) recommended by the system. In essence, the participant is choosing between asset configurations, because these are the configurations that

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Definition</th>
<th>Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeliness (T)</td>
<td>The probability that the asset will reach the new NAI in the time limit</td>
<td>Gain</td>
</tr>
<tr>
<td>Longevity (L)</td>
<td>The probability that the asset will be able to collect data for the required amount of time</td>
<td>Gain</td>
</tr>
<tr>
<td>Expected Coverage (EC)</td>
<td>The overall area that is expected to be covered by assets</td>
<td>Gain</td>
</tr>
<tr>
<td>Untimeliness (U)</td>
<td>The probability that the asset will arrive late at the new NAI</td>
<td>Loss</td>
</tr>
<tr>
<td>Disruptivity (D)</td>
<td>The probability that the asset will stop collecting data too soon because of battery or fuel problems</td>
<td>Loss</td>
</tr>
<tr>
<td>Expected Vulnerability (EV)</td>
<td>The overall area that is not covered by assets</td>
<td>Loss</td>
</tr>
</tbody>
</table>

Note: $U=1-T$, $D=1-L$, and $EV=1-EC$.|

<table>
<thead>
<tr>
<th>Configuarion No.</th>
<th>Levels of Three Asset Attributes</th>
<th>Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Timeliness</td>
<td>Longevity</td>
</tr>
<tr>
<td>C₁</td>
<td>High</td>
<td>Average</td>
</tr>
<tr>
<td>C₂</td>
<td>Average</td>
<td>High</td>
</tr>
<tr>
<td>C₃</td>
<td>Average</td>
<td>Average</td>
</tr>
<tr>
<td>C₄</td>
<td>Average</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>Untimeliness</td>
<td>Disruptivity</td>
</tr>
<tr>
<td>C₅</td>
<td>Low</td>
<td>Average</td>
</tr>
<tr>
<td>C₆</td>
<td>Average</td>
<td>Low</td>
</tr>
<tr>
<td>C₇</td>
<td>Average</td>
<td>Average</td>
</tr>
<tr>
<td>C₈</td>
<td>Average</td>
<td>Average</td>
</tr>
</tbody>
</table>

Note: Two values were established per attribute: average and high (or low in loss frame).
define an asset’s values. For instance, if ARV-1 is chosen, then the overall Timeliness, Longevity, and Expected Coverage of a configuration would be high, average, and average, respectively, and the configuration corresponding to these levels is $C_1$, as listed in Table III. Similarly, a configuration corresponding to choosing MULE-1 is $C_2$.

**Task #1 Procedure:** First, a participant was randomly assigned to either gain- or loss-framing condition and made a total of 18 pairwise comparisons among the four asset configurations (i.e., $C_1$–$C_4$ or $C_5$–$C_8$, as shown in Table III) with regard to the three asset attributes listed in Table II. In other words, if the participant was in the gain frame, he/she made six pairwise comparisons among the four configurations with regard to Timeliness, another six comparisons with regard to Longevity, and the other six comparisons with regard to Expected Coverage. A sample question follows.

**Question:** It has been determined that the following two assets will be the best choice to allocate to NAI X. With regard to Timeliness (or Untimeliness in the loss frame), which asset will you select?

Once the 18 comparisons are completed, each participant performed three more pairwise comparisons among the three attributes with regard to the overall objective. A sample question follows.

**Question:** Using a 9-point scale, compare the relative importance of the following two asset attributes with respect to the overall goal (i.e., select a better asset configuration). a) Timeliness versus Longevity (or Untimeliness versus Disruptivity in the loss frame).

To complete Task #1, the participant was then assigned to the other framing condition and made another 21 pairwise comparisons (18 comparisons for other frames, 4 configurations with regard to the three attributes, which had also been termed differently, and 3 comparisons for those three attributes).

**B. Task #2**

In Task #2, each participant answered pairwise comparison questions presented through a human-in-the-loop simulation, as shown in Fig. 4. The display showed a gain-framing condition and consisted of information available in Task #1 to include a map, a decision aid, an intelligence window, and a pairwise comparison question.

Compared with Task #1, Task #2 has four different characteristics: 1) The assets move from an initial position to a final one during each scenario; 2) the likelihood of an enemy presence and its size are shown in the right upper corner of the map and becomes more certain (i.e., approach unity) as a scenario progresses; 3) there was a time limit such that each simulation trial had to be completed in 20 s; and 4) in deciding between the two configurations, each participant needed to consider all three asset attributes simultaneously while the scenario progressed.

To measure judgment in this task, we did not stop the simulation to query each participant. Instead, we extracted preferences based on the values associated with each attribute on the configuration chosen, i.e., we asked which asset they would actually move, thereby forcing them to consider all attributes simultaneously.

**Task #2 Procedure:** As in Task #1, a participant was first randomly assigned to either framing condition and then answered a series of pairwise choice questions for the same candidate configurations used in Task #1. Once one frame is completed, the participant was assigned to the other frame and answered another series of pairwise choice questions for the four configurations, which had been framed differently.
In this task, the total number of pairwise choices was 12 to account for all possible pairwise combinations of the four asset configurations and the two different frames (i.e., six pairwise comparisons in the gain frame and the other six comparisons in the loss frame). In addition, each participant compared the three asset attributes with each other, in each frame, with respect to the overall objective for six more comparisons.

C. Task #3

Task #3 was similar in all aspects to Task #2 with one exception: The participant had only 10 s to view the decision aid and decide which asset to reallocate. This applied time pressure to the set of tasks. The Task #3 procedure was identical to the Task #2 procedure.

VI. MEASURES AND STATISTICAL ANALYSIS

Since the number of pairwise comparisons was different across the task types, we used outcome measures that resulted from the individual comparisons. Two dependent variables were calculated from the participants’ data using the AHP. One is the priority weights \( w_i \) (\( i = 1, 2, 3, 4 \)) of the four configurations shown in Table III under either framing condition for each task. The other is the CRH value under either framing condition for each task.

Due to the unit sum constraint on the priority weights (i.e., \( w_1 + w_2 + w_3 + w_4 = 1 \)), meaning that the priority weights are highly correlated, a traditional multivariate analysis of variance (ANOVA) (MANOVA) model could not be used to analyze the priority weights. To resolve this high-correlation issue (and, eventually, to perform MANOVA and then ANOVA on the priority weights), we used the c-compositional analysis \([34]–[37]\) and ln-ratio transformed the priority weights as follows:

\[
\begin{align*}
\text{ratio}_1 &= \ln \left( \frac{w_1}{w_4} \right) \\
\text{ratio}_2 &= \ln \left( \frac{w_2}{w_4} \right) \\
\text{ratio}_3 &= \ln \left( \frac{w_3}{w_4} \right)
\end{align*}
\]

We did this because an ln-ratio transformation of the compositional data makes the transformed variables independent, and statistical tests can then be applied \([34]\). In the transformation, we designated \( w_4 \), the priority weight of the fourth configuration, as the denominator in these ratios because the fourth configuration formed the smallest portion of the total weight (see Table IV), thus facilitating a better understanding of the relative contribution of the other configurations.

After the transformation, we checked the normality of the three ratios and found that they were very close to normal with zero mean and independence. Then, we proceeded to perform a two-way MANOVA test (shown below) on the three ratios.
To analyze the CRH values, we used a log transformation (base 10) on the CRH values because the variance across tasks was not constant. After transformation, we conducted an ANOVA on the transformed CRH values, $\log_{10}(CRH)$.

We note that because the focus of our work is on judgment performance rather than task performance, we did not examine task-performance outcomes such as latency and accuracy.

### VII. Results

For all statistical tests, an alpha level of 0.05 was used.

$H_1$ and $H_2$: Weight prioritization was analyzed by conducting a MANOVA model on the ln-ratio-transformed priority weights of the four alternatives for the task type (i.e., static, human-in-the-loop simulation, and human-in-the-loop simulation under time constraint) and framing (i.e., gain and loss) considered. Since the MANOVA results showed the significant effects of framing (Wilks’ $\Lambda = 0.376$, $F_{(3, 45)} = 24.86$, and $p < 0.001$), task type (Wilks’ $\Lambda = 0.396$, $F_{(6, 184)} = 18.06$, and $p < 0.001$), and the interaction of task type and framing (Wilks’ $\Lambda = 0.689$, $F_{(6, 184)} = 6.28$, and $p < 0.001$), we proceeded to conduct separate univariate ANOVA tests on the three ratios (Table V) and found that there was a significant difference between gain and loss frames for all ratios except $\text{ratio}_2$ and that the task type effect was also significant for all ratios. In addition, the interaction (frame $\times$ task type) effect was significant such that, for all three ratios, the difference between gain and loss frames was greater in Task #1 than in those in Tasks #2 and #3.

To further examine which priority weight has been changed significantly, we conducted univariate ANOVA tests on the ln-transformed priority weights (Table VI) and found that there was a significant difference between gain and loss frames for all ln-transformed weights except $\ln(w_1)$. We also found that the task type effect was also significant for all ln-transformed weights except $\ln(w_2)$ and that, for the interaction (framing $\times$ task type) effect (see Fig. 5), only $\ln(w_2)$ and $\ln(w_4)$ showed significant results such that, for $\ln(w_2)$, which is the weight given to $C_2$, the difference between the gain and loss frames was increased as the complexity of task increased. In addition, for $\ln(w_4)$, participants gave considerably different weights to $C_4$ under different frames in Task #1, whereas in Tasks #2 and #3, they gave similar weights to $C_4$ under either framing conditions.

Through Tukey’s post hoc tests shown in Table VII, it was also found that the participants gave significantly different weights under different frames and also under different task conditions. Specifically, participants gave significantly lower weights to $C_2$ and $C_4$ and higher weight to $C_3$ under gain frame.
TABLE VII
TUKEY’S POST HOC COMPARISONS OF ln-RATIO TRANSFORMED WEIGHTS FOLLOWING SIGNIFICANT UNIVARIATE TESTS

<table>
<thead>
<tr>
<th></th>
<th>Task #1</th>
<th>Task #2</th>
<th>Task #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(w1)</td>
<td>-1.305(0.661)a</td>
<td>-1.406(0.629)a</td>
<td></td>
</tr>
<tr>
<td>ln(w2)</td>
<td>-1.715(0.844)a</td>
<td>-1.309(0.629)b</td>
<td></td>
</tr>
<tr>
<td>ln(w3)</td>
<td>-1.147(0.583)a</td>
<td>-1.326(0.591)b</td>
<td></td>
</tr>
<tr>
<td>ln(w4)</td>
<td>-3.150(0.843)a</td>
<td>-2.808(1.141)b</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values are given as the mean, with SD in parentheses. Sample sizes are 48 in all cases. Further, higher levels of CRH indicate lower consistency.

TABLE VIII
CONSISTENCY RATIO OF THE HIERARCHY (CRH) VALUES

<table>
<thead>
<tr>
<th></th>
<th>CRH</th>
<th>Task #1</th>
<th>Task #2</th>
<th>Task #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain</td>
<td>0.186(0.157)</td>
<td>0.952(1.047)</td>
<td>1.642(1.323)</td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td>0.361(0.304)</td>
<td>1.473(1.053)</td>
<td>2.382(1.327)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values are given as the mean, with SD in parentheses. Sample sizes are 48 in all cases.

Fig. 5. Interaction plots for (a) ln(w2) and (b) ln(w4).

VIII. DISCUSSION

The results conform to findings in traditional judgment literature, which rely on utility theoretic axioms. However, instead of deviation from an expected value, we offer an alternative explanation in terms of judgment consistency between and within treatment conditions subject to framing and time-pressure effects.

A. Impact of Framing and Task Type on Weight Prioritization (H₁ and H₂)

Our results suggested that there exists a statistical effect of framing and task type on the weight prioritization of sensor configurations. The mean values of ln-ratio-transformed weights and ln-transformed weights show that participants weighed the same alternatives differently under different frames and in different task conditions. Our findings are consistent with the argument that dynamic decision environments impact judgment behavior [38]. Specifically, our results are consistent with findings in decision research that time pressure alters one’s choice behavior through inducing suboptimal judgment heuristics [39]–[41].

B. Impact of Framing on Judgment Consistency (H₃)

H₃ and H₄ address the internal consistency of the decision behavior under framing and time-pressure conditions. Our findings suggest that CRH values represent the stability of weight prioritization across treatments and can be used as the potential first step for establishing a coherence criterion for truth [17]. Although utility theory has served as such a criterion for internal consistency, we propose the use of CRH as an alternative criterion to assess judgment consistency of participants. Specifically, we found judgment consistency to significantly worsen under loss compared with gain framing. Hence, it appears that the weight prioritization under loss framing is more volatile than that under gain framing. When
information is presented negatively in the loss frame, participants’ judgments about the relative importance of different attributes were less likely to follow consistent internal rules. We think that this finding complements the stipulation of Prospect Theory [11] that the value function for losses has a different slope than the corresponding function for gains. However, we also acknowledge that further investigation is needed to confirm our notions.

C. Impact of Task Type on Judgment Consistency ($H_4$)

We found consistency to deteriorate as the time available to choose a configuration decreases. Like our findings on weight prioritization, our results on judgment consistency are consistent with findings that time pressure [40], [41] affects choice behavior. Thus, our findings regarding the effect of time pressure are consistent with what would be expected if participants had been using cognitively less demanding strategies. While we do not have access to the heuristics used by the participants, we do know that commonly acknowledged heuristics such as Take-the-Best [17], Elimination-by-Aspects [42], and Lexicographic [43] do not guarantee transitivity in preferences over multiple attributes.

In general, participants made significantly different choices under the gain versus loss frames under time pressure. Perhaps the cause of the discrepancy is due to the formation of different judgment heuristics under time pressure. Interestingly, in spite of the uneven number of pairwise comparisons between Task #1 and Tasks #2 and #3, the effect of time pressure (Task #3) was still significant. Nevertheless, we recommend that further investigation of the impact of uneven comparisons be made.

IX. CONCLUSION

In this paper, we have developed a methodology to assess judgment performance in a real-time multiattribute judgment task. Specifically, we used the AHP to conduct pairwise comparisons to evaluate human judgment consistency under framing and time-pressure effects. We found that framing significantly affected the prioritization of asset configurations. Moreover, we found judgment consistency to significantly worsen under the loss frame. With regard to time pressure, we found that judgment consistency deteriorates as the time to choose an asset configuration decreases.

This paper contributes to judgment literature by providing an evaluation of the impact of framing and task type on judgment performance. Through an understanding of the impact of task conditions on configuration prioritization, we gained an ability to assess the limitations of judgment performance. Additionally, through an understanding of the effects of task conditions on judgment consistency, we can analyze the reliability of human judgment.

This paper differs from utility-theoretic frameworks like Prospect Theory in that we can assess judgment consistency without requiring the use of task-performance outcomes. Moreover, we believe that this work is a step toward establishing a coherence criterion [17] through the use of the CRH to investigate judgment under naturalistic conditions. This is particularly helpful in complex envisioned domains such as robotic asset allocation tasks in which tactics and procedures have not yet been developed and researchers cannot define the one “right” answer. In the absence of a right answer, the CRH provides a tool that system developers can use to determine if the answers given using specific displays are internally consistent. An additional consideration, which we would like to investigate in the future, is the impact of framing and task type on subject-matter experts. In fact, we think that the tools offered by the AHP can help us to better measure expert consistency.

APPENDIX

EXAMPLE OF CALCULATING PRIORITY WEIGHTS AND CONSISTENCY RATIO OF THE HIERARCHY

Suppose that a participant generated the reciprocal matrices (Fig. 6) for our problem, which is to select a better asset configuration.

To calculate the overall priority weights of the four configurations, we first need to calculate the weight vector $w$ of the four matrices by normalizing the principle eigenvector of the corresponding matrix. Once calculated, the $w$ for the four matrices (i.e., one for the three asset attributes and the other
three for the four configurations with respect to each attribute) would be

\[
\begin{bmatrix}
0.701 & 0.604 & 0.631 & 0.127 \\
0.193 & 0.213 & 0.182 & 0.281 \\
0.106 & 0.064 & 0.070 & 0.120 \\
0.119 & 0.117 & 0.463 & 
\end{bmatrix}.
\]

In case of a multilevel hierarchy (like ours), these weight vectors are aggregated into a performance matrix \(W^h\) for each level \(h\) of the hierarchy (see Fig. 2) for \(h = 2\) to \(v\) (in our case, \(v = 3\)) as follows:

\[
W^2 = \begin{bmatrix} 0.701 \\ 0.193 \\ 0.106 \end{bmatrix} \quad W^3 = \begin{bmatrix} 0.604 & 0.631 & 0.127 \\ 0.213 & 0.182 & 0.281 \\ 0.064 & 0.070 & 0.120 \\ 0.119 & 0.117 & 0.463 \end{bmatrix}.
\]

In this case, \(W^2\) reflects the relative weights of the three attributes with respect to the objective, while \(W^3\) represents a composite of the weights generated by comparing the four configurations with respect to the three asset attributes.

Finally, using (3), the priority weights of the four configurations with respect to the overall objective \(U\) is calculated as

\[
U = W^3W^2 = \begin{bmatrix} 0.559 \\ 0.214 \\ 0.071 \\ 0.155 \end{bmatrix}.
\]

To calculate the consistency ratio of the hierarchy \((CRH)\), we use (4)–(8). For example, for the reciprocal matrix of the three assets attributes, \(CI\) and \(CR\) are calculated as follows:

\[
CI = \frac{\lambda_{\text{max}} - n}{n - 1} - 1 = 0.129
\]

\[
CR = \frac{CI}{RI} = \frac{0.129}{0.58} = 0.222
\]

where \(\lambda_{\text{max}}\) is the largest eigenvalue of the reciprocal matrix (i.e., 3.257) and \(RI\) for \(n = 3\) is 0.58 (see Table I).

The calculated \(CI\) and \(RI\) values for the other reciprocal matrices are shown right below the corresponding matrices.

Finally, the \(CRH\) value is calculated as follows:

\[
M = CI^2 + W^2 \begin{bmatrix} CI^2 \\ CI \\ CT \end{bmatrix}
\]

\[
= 0.129 + \begin{bmatrix} 0.701 \\ 0.193 \\ 0.106 \end{bmatrix} \begin{bmatrix} 0.116 \\ 0.148 \\ 0.473 \end{bmatrix} = 0.289
\]

and if we take \(RI^2 = 0.58\) and \(RI^3 = 0.90\) from Table I

\[
\tilde{M} = RI^2 + W^2 \begin{bmatrix} RI^3 \\ RI \end{bmatrix}
\]

\[
= 0.580 + \begin{bmatrix} 0.701 \\ 0.193 \\ 0.106 \end{bmatrix} \begin{bmatrix} 0.900 \\ 0.900 \end{bmatrix} = 1.480.
\]

Hence

\[
CRH = \frac{M}{\tilde{M}} = \frac{0.289}{1.480} = 0.195.
\]

**References**


[28] E. E. Spires, “Using the analytic hierarchy process to analyze multi-
attribute decisions,” Multivariate Behav. Res., vol. 26, no. 2, pp. 345–361,
[31] T. L. Saaty, “An exposition on the AHP in reply to the paper ‘remarks on
the analytic hierarchy process’,” Manage. Sci., vol. 36, no. 3, pp. 259–268,
[33] P. T. Harker and L. G. Vargas, “The theory of ratio scale estima-
tion: Saaty’s analytic hierarchy process,” Manage. Sci., vol. 33, no. 11,
analysis of habitat use from radio-tracking data,” Ecology, vol. 74, no. 5,
behavior at spring-migration stopover areas in Southern Alberta,” Can. J.
[37] J. L. Beck and J. M. Peek, “Diet composition, forage selection, and
potential for forage competition among elk, deer, and livestock on
aspen–sagebrush summer range,” Rangeland Ecol. Manage., vol. 58,
no. 2, pp. 135–147, 2005.
[38] R. Lipshitz, “Converging themes in the study of decision making in real-
istic settings,” in Decision Making in Action: Models and Methods, G. A.
Klein, J. Orasanu, R. Calderwood, and C. Zsambok, Eds. Norwood, NJ:
Ablex, 1993, pp. 103–137.
Time Pressure and Stress in Human Judgment and Decision Making,
in decision making,” J. Exp. Psychol.: Learn., Memory, Cognition, vol. 14,
[41] P. Wright, “The harassed decision maker: Time pressures, distractions,
and the use of evidence,” J. Appl. Psychol., vol. 59, no. 5, pp. 555–561,
1974.
maker: Effort and accuracy in choice,” in Insights in Decision Making:
A Tribute to Hillel J. Einhorn, R. M. Hogarth, Ed. Chicago, IL: Univ.

Jeonghwan Jin received the B.S. degree in industrial engineering from Jeonju University, Jeonju, Korea, in 1996 and the M.S. degree in manufacturing system engineering from the University of Warwick, Coventry, U.K., in 1998. He is currently working toward the Ph.D. degree in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park. His research interests include human behavior modeling, human–robot interaction, and decision-making processes.

Ling Rothrock (SM’06) received the Ph.D. degree in industrial engineering from the Georgia Institute of Technology, Atlanta, in 1995.

He was an Officer with the U.S. Army until 2000. He is currently an Associate Professor with the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park. His research interests include human-in-the-loop simulations, behavioral decision making, display visualization, and human–machine performance assessment. He has published over 60 research articles in these areas. Dr. Rothrock is also a member of the Institute of Industrial Engineers.

Patricia L. McDermott received the M.S. degree in human factors engineering from Wright State University, Dayton, OH, in 2001.

She has over ten years of experience in the field of naturalistic decision making and human factors and is currently with Alion Science and Technology, Boulder, CO. Her research interests include human–robot interaction and decision making under uncertainty.

Michael Barnes received the M.S. degree in experimental psychology from New Mexico State University, Las Cruces.

He is currently with the U.S. Army Research Laboratory, Fort Huachuca, AZ, where he is the human–robot interaction Lead for the Army Technology Objective in Safe Operations in Unmanned Reconnaissance missions in Complex Environments. He was a Researcher with the Naval Air Warfare Center Weapons Center, China Lake, CA, and the Naval Air Development Center, Warminster, PA, and was also the Aegis Combat System human factor’s Unit Manager with the Government Services Division, General Electric Cooperation. He is currently the Chief of the Fort Huachuca field element where he supports cognitive research in intelligence analysis, unmanned aerial vehicles, and visualization tools for risk management. He has published over 70 research articles.