2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Recency Bias in Task Performance History Affects Perceptions of Robot Competence and Trustworthiness

Matthew B. Luebbers*, Aaquib Tabrez*, Kanaka Samagna Talanki, and Bradley Hayes

***Abstract*— Human memory of a robot's competence, and resulting subjective perceptions of that robot, are influenced by numerous cognitive biases. One class of cognitive bias deals with the ordering of items or interactions: information presented last among a grouping is most salient in memory formation (recency bias), followed by information presented first (primacy bias), followed by information in the middle, collectively known as the serial-position effect. For example, if a human's last observation of a robot involves a task failure, this will disproportionately negatively alter their perception of the robot's competence, as well as their trust in the robot moving forward. It is valuable to characterize the effect of these biases within human-robot interactions to inform strategies for risk-aware planning that cultivate appropriate levels of human trust. We conducted a human-subjects study (n=53) testing the influence of the serial-position effect on recalled competence (see overview at https://youtu.be/BgH2zhh1s48). Participants viewed videos of a robot performing the same tasks at the same level of competence, with task order differing by experimental condition (rising competence, falling competence, or failures at the midpoint), asking participants to rate robot competence in between every video as well at the very end of the experiment. We found that while the average between-video rating of robot competence remained stable across conditions, the recalled, post-experiment ratings of competence and trust were significantly lower in the condition with decreasing competence than in either of the other two conditions, suggesting a notable recency bias. We conclude with implications for human-subjects experiment design (i.e., how subjective measures are influenced by ordering effects) and provide design recommendations to minimize them. We further discuss practical applications of these results in creating risk-aware robotic planners capable of trust calibration.**

I. INTRODUCTION

Human recall of events and experiences is subject to the influence of various cognitive biases. Specifically, individuals have a tendency to simplify their memories, selectively retrieving significant aspects of an interaction to make snap judgments, in accordance with mental shortcuts like the representativeness heuristic [14]. The particular behavioral patterns that stem from this process are known as memory biases. One such bias is the serial-position effect [3], according to which individuals tend to systematically recall the latest items or interactions in a series the best (recency bias), followed by the earliest items or interactions (primacy bias), followed by all interactions in the middle.

In this work, we hypothesize that human perception of a robot's competence is strongly influenced by the ordering of their interactions with that robot, in accordance with the relative importance of late, early, and middle interactions indicated by the serial-position effect. That is, when individuals interact with a robot performing the same tasks, at the same level of competence, the order in which sequences of robot success and failure occur will alter their post-hoc rating of robot competence, with what they saw at the end having the biggest impact, followed by what they saw at the beginning.

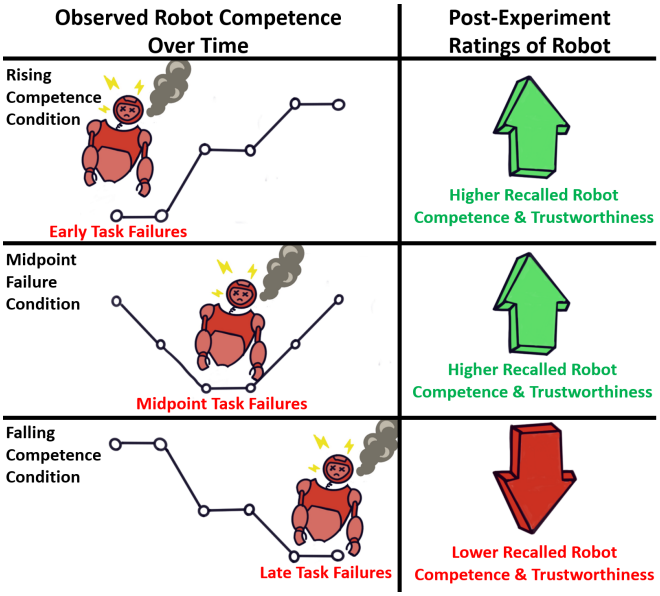

Fig. 1: We conducted an online human-subjects study where participants viewed videos of a robot performing the same tasks at the same level of competence in three distinct orderings, depending on their assigned condition: rising competence (involving early task failure), midpoint failure, and falling competence (with late failure). We found that participants recalled the robot as being significantly less competent and trustworthy in the falling competence condition compared to the other two conditions, indicating the presence of a recency bias on those subjective perceptions.

An individual's perception of a robot holds implications not only for their immediate interaction with that robot, but also for its usage and deployment in future scenarios [18]. We posit that memory biases would not only impact a human's recalled rating of robot competence, but will also impact related measures of the robot's trustworthiness. If a robot is perceived as incompetent and error-prone, it is less likely to be trusted in the future, potentially leading to negative outcomes in critical situations [17]. Conversely, if a robot is perceived as highly competent and capable, collaborators are

*These authors contributed equally to this work.

This work was funded by the Army Research Lab STRONG Program (#W911NF-20-2-0083).

The authors are affiliated with the Department of Computer Science, University of Colorado Boulder, 1111 Engineering Drive, Boulder, CO 80309, U.S.A. {firstname}.{lastname}@colorado.edu

more inclined to rely on it, leading to increased utilization moving forward. This highlights the practical implications of interaction ordering - even if a robot's level of competence remains consistent on average, the timing of robot activities more or less prone to failure will significantly alter trust in that robot going forward if they are placed at the beginning or end of the interaction.

This phenomenon has implications for how user studies are conducted within human-robot interaction (HRI) research. Many user studies rely on post-hoc subjective measures to assess participants' perceptions of robot performance and trustworthiness, which have the potential to be influenced by memory biases. It is therefore important to take these biases into account when designing user studies aimed at measuring human perceptions of robots, and take appropriate counter-measures if necessary. We present the following contributions:

- A human-subjects study design to test for the existence of recency and primacy effects on participant-recalled ratings of robot competence and trustworthiness.
- Analysis with statistical evidence showing that task failures at the end of an interaction result in lower post-experiment ratings of competence and trust, despite the robot having performed the same tasks at the same overall level of competence as other conditions, indicating the presence of a recency bias (Fig. 1).
- A set of HRI user study design strategies for mitigating the effect of ordering biases in subjective measures.
- Strategies to incorporate knowledge of ordering biases into robot planners to deliberately calibrate a human's trust in a robot.

II. RELATED WORK

Cognitive and Memory Biases in Psychology. Humans rely on a variety of mental shortcuts, known as heuristics, to ease the cognitive load of decision-making. Although these heuristics are not always rational or optimal, they assist individuals in problem-solving and forming judgments about events. Herbert Simon first introduced the concept of 'satisficing', wherein people accept choices, solutions, and judgments that meet an acceptable quality threshold for their practical purposes, despite their suboptimality [27].

Similarly, recall of past events and experiences is influenced by processes of mental simplification, governed by known patterns hard-wired by evolution and measurable through experiment, known as memory biases [15]. Memory biases selectively enhance or impair the recall of specific memories, including the likelihood of the memory being recalled at all, the time it takes to recall it, or the degree to which it is altered when recalled [31]. Many of these biases were discovered through experiments involving free recall tasks, wherein participants study a list of items and are prompted to recall the items in any order [23]. One such memory bias is the serial-position effect, wherein people tend to recall the last items of a series or list the best, followed by the first items of a series, and lastly, items in the middle [3]. These sub-effects are known as the recency bias (the tendency to disproportionately recall the last items of a list), and the primacy bias (the tendency to disproportionately recall the first items of a list).

This effect has also been examined in persuasive communication studies, where the order of information presentation influences the formation of opinions [16]. Smith et al. tested the impact of ordering effects on the assessment of sports ability [28]. Participants watched a video of an ultimate Frisbee player demonstrating skills presented in ascending or descending order of ability, and made assessments of the player's overall ability. The researchers found the presence of a primacy effect when assessments were made at the end of the video, but not when assessments were made after watching each skill. Similarly, Garnefeld and Steinhoff conducted a study on the impact of ordering effects on customer satisfaction with hypothetical hotel stays [8]. Participants received daily descriptions of their stay with differing sequences of positive, negative, and neutral experiences. The study found that the timing of positive and negative events had an effect on customer satisfaction, with a recency effect for negative events and a primacy effect for positive events.

Cognitive biases such as these have been extensively applied in marketing, UI/UX design, health care, customer service, and policy recommendation [29]. They are leveraged in UI/UX design for creating easy to use interfaces and improving engagement metrics [33], or in medicine for designing patient interactions during potentially painful procedures that are remembered more favorably after the fact [19]. Closely related to the serial-position effect, a phenomenon known as the peak-end effect is used prominently for this purpose, in which memories of experiences are disproportionately influenced by the most intense point (the "peak") and the end of the experience [7]. Deliberately reordering certain painful or painless portions of medical procedures like colonoscopies according to the peak-end effect has been shown to lead patients to rate the entire experience as being less painful overall [13].

Cognitive Biases and Human-Robot Interaction. Cognitive biases also influence how humans interact with robots. For instance, anthropomorphism bias is often seen in HRI, with participants rating robots with human-like characteristics more favorably [24]. Similarly, people prefer to render assistance to robots that are similar in appearance and perceived personality to themselves, rather than ones best suited for that task [30]. Framing bias has been demonstrated in language-enabled robots, with positive framing in robotic communication leading to more favorable perceptions of the robot [25]. There is also some evidence of memory bias in HRI, with first impressions found to significantly influence trust in a robot's advice, even after observing the robot malfunctioning later in the interaction [34].

Memory biases have significant implications for how user studies are conducted in HRI research. Desai et al. investigated the impact of changing reliability on the trust of robot systems and control allocation strategies [5]. The study found that trust was affected by drops in robot reliability, with mode-switching behavior (voluntarily taking control of

the robot) becoming more frequent as reliability dropped. Though the study showed slightly higher likelihood for mode-switching when the reliability dropped in the middle or later stages of runs compared with the beginning, the effect did not have statistical power. The authors extended the work further, comparing conventional post-run trust survey approaches to periodic polling on trust at regular intervals, showing that live trust falls precipitously following drops in reliability, and slowly climbs as the robot performs well [4]. Our work aims to isolate the measurable effects of ordering biases as they relate to impressions of a robot's competence, and resultant trust in the robot. Furthermore, our research offers techniques to both mitigate these effects when designing user studies in HRI in order to obtain more accurate results, as well as leverage them for trust calibration in practical HRI scenarios.

III. METHODOLOGY

We conducted an IRB-approved, online human-subjects study to investigate the role of ordering biases on perception of robot competence and trustworthiness in HRI scenarios.

A. Experimental Task: Kitting

For our experimental evaluation, we adopted a robotic kitting task similar to the task presented by Tung et al. [32]. While traditional kitting involves preparing and grouping necessary parts and tools into predefined collections or "kits" for later assembly in a manufacturing environment, autonomous kitting allows for additional flexibility in assembly while keeping storage space requirements low and worker productivity high [32]. We chose this task because it satisfies two design criteria. First, it is a validated application with relevance to the human-robot interaction community. Second, it is important for participants to be able to easily comprehend robot success or failure by watching it conduct its tasks, forming judgments about its competence.

Robotic Platform. For our experiment, we used a Sawyer robotic manufacturing arm. The Sawyer robot was tasked with aiding an unseen human worker in the assembly of a small flat-pack furniture table by delivering the required parts to the human. Specifically, Sawyer attempted to pick up various parts and transport them to the target bin, forming kits for delivery to the human. These delivery tasks were filmed and shown to participants as videos to keep the robot competence consistent between trials.

Table Assembly. We divided the entire delivery process needed to assemble the flat-pack furniture table into six sub-tasks, as outlined in Table I. Each sub-task was performed by the Sawyer robot at varying levels of competence, and the videos of the sub-tasks were ordered according to the experimental condition. We defined three levels of competence for the robot: (1) Success, in which all required parts for the task were successfully retrieved and delivered; (2) Partial Success, in which a subset of the required parts were retrieved and delivered; and (3) Failure, in which none of the required parts were retrieved and delivered.

B. Experimental Design

To evaluate the presence of ordering biases in perceptions of robot competence and trustworthiness, we conducted an online video study using the Prolific platform (prolific.co). The study employed a 3x1 between-subjects design, in which participants viewed six different videos of the Sawyer robot performing individual sub-tasks (attempting to pick up specified parts from a table and deliver them to a target bin, shown in Fig. 2). Two of the sub-task videos were designed to be successful, two partially successful, and two failing. Participants were randomly assigned to one of three conditions, determining how the six videos were ordered:

- **Rising Competence Condition:** The 6 sub-task videos were ordered in ascending order of competence, starting with two failures (F1, F2), then two partially successful tasks (P1, P2), and ending with two fully successful tasks (S1, S2); *(order: F1, F2, P1, P2, S1, S2)*.
- **Midpoint Failure Condition:** The 6 sub-task videos were ordered with a negative peak in the middle, starting with a successful task, followed by a sharp decrease to failure and then a sharp increase back to a successful ending task; *(order: S1, P1, F1, F2, P2, S2)*.
- **Falling Competence Condition:** The 6 sub-task videos were ordered in descending order of competence, starting with two fully successful tasks, then two partially successful tasks, and ending with two failed tasks; *(order: S1, S2, P1, P2, F1, F2)*.

Fig. 2: Left: the experimental task setup. The Sawyer robot attempts to move construction components from the left of the table into the blue bin on the right, pushing the bin forward to deliver it to the human. Right: the components Sawyer can deliver as part of its sub-tasks.

C. Hypotheses

Isolating the individual components of the serial-position effect, we evaluate the following hypotheses regarding hu-

Sub-Task	Components Required	Components Retrieved	Competence Level
1	1x Dowel + 1x Small Screws	1x Dowel + 1x Small Screws	Success (S1)
2	3x Foot	3x Foot	Success (S2)
3	1x Top + 1x Large Screws	1x Large Screws	Partial Success (P1)
4	3x Top	2x Top	Partial Success (P2)
5	3x Dowel	Nothing	Failure (F1)
6	1x Foot + 1x Nuts	Nothing	Failure (F2)

TABLE I: Description of experimental sub-task videos. The ordering of videos 1 - 6 varies by experimental condition.

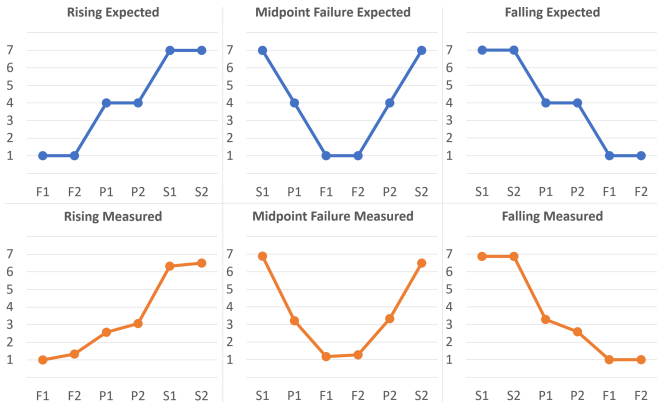

Fig. 3: Top: Expected patterns of rated robot competence per sub-task per condition, consisting of successes (7 on a 1-7 Likert scale), failures (1 on a 1-7 Likert scale), and partial successes with an intermediate rating. Bottom: Measured ratings of robot competence per sub-task per condition.

man recalled perception of robot competence and trustworthiness with our experiment:

- **H1: Recency Bias Hypothesis**: Participants' recalled perception of robot competence and trustworthiness will be lower in the 'Falling Competence' condition than in the 'Midpoint Failure' condition. This is because, though both conditions start the same, negative events at the end of an interaction are likely to have an outsized effect on memory.
- **H2: Primacy Bias Hypothesis**: Participants' recalled perception of robot competence and trustworthiness will be lower in the 'Rising Competence' condition than in the 'Midpoint Failure' condition. This is because, though both conditions end the same, negative events at the start of an interaction are likely to have an outsized effect on memory.
- **H3: Serial-Position Hypothesis**: Participants' recalled perception of robot competence and trustworthiness will be lower in the 'Falling Competence' condition than in the 'Rising Competence' condition. This is because, within the serial-position effect, recency bias is stronger than primacy bias.

D. Study Protocol

The experiment was administered online in several batches with randomly assigned conditions using the Prolific platform to crowdsource participants. To ensure higher participant quality, we filtered for those who had both completed at least 100 approved studies on Prolific and possessed an approval rate of 95% or higher. Upon providing informed consent, participants were given a description of the Sawyer robot and the kitting task, with pictures of all the required parts for the Sawyer to deliver. Participants then watched a series of six videos, each showcasing an individual subtask (see Table I), with order depending on their assigned experimental condition. After each video, they were asked a comprehension question related to the video to assess attention, followed by a post-video survey gauging impressions of robot competence in the video they just watched.

After watching all six videos and completing the post-video surveys, participants completed a brief memory-matching distractor task to provide a clear boundary between the experimental task as a whole (and the associated between-video questions) and the final survey asking for post-hoc recalled perceptions of the robot, moving the experiment from active memory to recall. The final survey consisted of a combination of Likert scale, free response, and demographic questions. The Likert scale questions were designed to measure participants' attitudes and perceptions towards the Sawyer robot after the entire interaction had concluded, a common subjective data collection practice in human-robot interaction research.

E. Measurement

We recruited 54 participants (25 male, 24 female, 4 nonbinary, 1 non-specified) through the Prolific platform, with an age range of 19 to 69 years old (M = 38.64, SD = 14.80). 15% of participants reported working in a STEM field, and 51% of participants reported having received a bachelor's degree or higher. Although we gathered data from all participants (18 per condition), one participant's data was removed from analysis as they failed both the video comprehension check and survey attention check questions.

We utilized multiple subjective measures to evaluate our hypotheses. We administered two sets of questionnaires: short, post-task questionnaires after each video asking individual questions about robot competence, and a longer, post-experiment questionnaire administered after watching every video, containing questions corresponding to the post-task questionnaire items, as well as additional sets of questions developed using established scales from HRI literature, such as the Trust in Automation Survey [12], Trust Perception Scale-HRI [26], Positive Teammate Traits [10], and Robotic Social Attributes Scale (RoSAS) [1]. Based on the responses to these questionnaires, we identified two relevant concepts to validate our hypotheses: *Competence* and *Trust*.

To isolate these concepts, we conducted a principal component analysis using the scales mentioned and calculated the factor loading matrix using varimax rotation. We selected items that could be combined to create concept scales with a correlation cutoff of $r \geq 0.6$ to the factor matrix [11]. The resulting scales are presented in Table II.

TABLE II: Subjective scale measure items.

Competence (Cronbach's α = 0.92)
1. Sawyer was competent at completing its tasks.
2. I can rely on Sawyer to correctly perform its tasks.
3. Sawyer was efficient at performing its tasks.
4. I feel confident that Sawyer is competent at performing its tasks.
5. Sawyer was dependable.
6. Sawyer was incompetent in performing its tasks. [inverted scale]
7. Sawyer failed to complete its tasks regularly. [inverted scale]
Trust (Cronbach's α = 0.82)
1. I trust that Sawyer will perform its tasks successfully.
2. Sawyer was trustworthy.
3. Sawyer was committed to its tasks.
4. Sawyer was capable of performing its tasks successfully.
Likert items are coded as 1 (Strongly Disagree) to 7 (Strongly Agree)

IV. RESULTS

We analyzed the between-video and post-experimental survey responses to test our hypotheses. For comparisons of results between conditions, we utilized a one-way analysis of variance (ANOVA) with experimental condition as a fixed effect. Post-hoc tests used Tukey's HSD to control for Type I errors in comparing results across conditions.

The ANOVA revealed a significant effect for the post-experimental *Competence* scale described in Section III ($F(2,50) = 4.24$, $p = 0.020$). Post-hoc analysis with Tukey's HSD revealed that participants recalled the robot as being significantly less competent in the falling competence condition ($M = 2.01$) compared with both the rising competence ($M = 2.93$), $p = 0.032$ and midpoint failure conditions ($M = 2.88$), $p = 0.045$ (Fig. 4 left). This indicates the presence of a recency bias, since the condition with low performance at the end of the interaction was recalled as being more competent than either condition with high performance at the end.

Significant differences were identified from the ANOVA comparing conditions on the *Trust* scale ($F(2,50) = 4.62$, $p = 0.014$). Post-hoc analysis showed the same pattern, with participants trusting the robot less post-experiment in the falling competence condition ($M = 2.59$) compared with both the rising competence ($M = 3.68$), $p = 0.027$, and midpoint failure conditions ($M = 3.65$), $p = 0.032$ (Fig. 4 right). This highlights the correlation of recalled robot competence with trust in the robot's ability to perform looking forward.

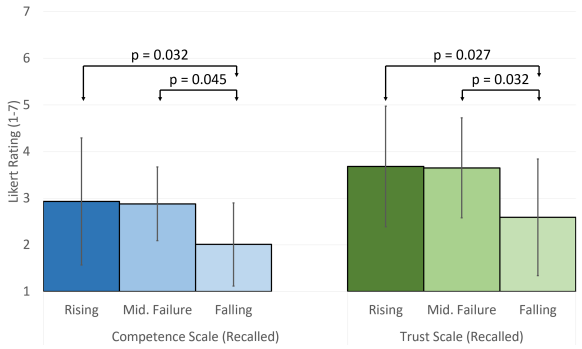

Fig. 4: Average scores per condition for the post-experimental *Competence* (left), and *Trust* (right) scales.

Since recalled robot competence and trustworthiness were both higher in the midpoint failure condition compared to the falling competence condition, we can **validate Hypothesis 1** and demonstrate the existence of a recency bias in these recalled measures. Since the same measures yielded no significance for the midpoint failure condition compared to rising, **Hypothesis 2 is inconclusive** with no direct evidence found showing a primacy effect. Lastly, since the rising condition also outperformed the falling condition, we can additionally **validate Hypothesis 3**, showing the observed recency bias to be stronger than any potential primacy bias.

We also analyzed the difference between participants' average response to the six between-video Likert-scale questions stating "I think Sawyer performed its task competently in the video I just watched," and the post-experimental Likert-scale question stating "Sawyer was competent at completing its tasks." Within each condition, for comparing between-video competence and post-experiment (recalled) competence, we utilized a one-tailed t-test, testing whether the recalled competence rating is significantly lower than the average of between-video competence ratings.

Showing alignment with our other findings, we found that only the falling competence condition showed a significant dropoff, with the average between-video rating for falling ($M = 3.61$), being significantly higher than the average recalled, post-experiment rating for falling ($M = 2.18$), $p < 0.0001$. No effect was found between the average between-video rating for midpoint failure ($M = 3.73$) and its recalled rating ($M = 3.56$), or the average between-video rating for rising competence ($M = 3.46$) and its recalled rating ($M = 3.44$), as shown in Fig. 5. This reinforces evidence of a recency bias.

We additionally compared the average between-video and recalled ratings of competence independently using one way ANOVAs, with experimental condition as a fixed effect. Comparing the between-video ratings of each condition yielded no significant differences. Comparing the recalled ratings, however, did show significant differences ($F(2,50) = 6.93$, $p = 0.0022$). Post-hoc analysis shows that, similarly to the competence and trust scales shown in Fig. 4, scores on the post-experimental question "Sawyer was competent at completing its tasks." were significantly lower in the falling competence condition ($M = 2.18$) compared with both the rising competence ($M = 3.44$), $p = 0.0089$, and the midpoint failure conditions ($M = 3.56$), $p = 0.0041$. These results are also shown in Fig. 5.

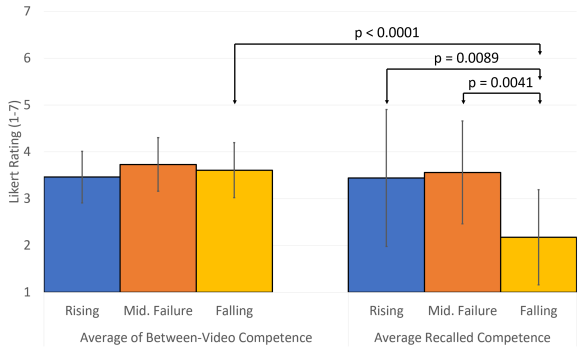

Fig. 5: Average scores per condition for the six between-video questions asking "I think Saywer performed its task competenty in the video I just watched" (left), and average scores for the post-experimental question "Sawyer was competent in completing its tasks" (right).

Result Synopsis: These results serve to reinforce the evidence for a recency bias in perceptions of robot competence. They also demonstrate that obtaining this measure at regular intervals rather than at the end of an interaction can counteract that recency bias, as the three conditions led to similar between-video rated competence, an objectively more accurate result than the recalled competence scores since the robot performed the same tasks in each condition at the same level of competence (see https://youtu.be/BgH2zhh1s48).

V. DISCUSSION AND TAKEAWAYS

We describe the implications of our experimental findings for HRI regarding the serial-position effect across two broad categories: experimental design and practical applications.

A. Experimental Design Recommendations

Our findings on ordering effects have practical implications for the design of HRI user studies. Here, we summarize key takeaways from our results that researchers should consider when designing HRI experiments, discuss the potential consequences of overlooking these effects, and provide recommendations for mitigating them.

Recency Bias: Due to recency bias, negative interactions that occur towards the end of an experiment will lead to an outsized negative impact in participant perception and subjective ratings of a robot across multiple measures. For example, if a study requires a robot to experience failures at some point, those failures occurring at the end would likely overload post-experimental subjective measures. To mitigate this effect, care should be taken to deliberately extend experimental interactions beyond such failures, ensuring that the interaction does not end on an atypical note (as is done in medical procedures like colonoscopy or lithotripsy [13]). Although our study did not confirm a primacy effect, related research has shown that positive first impressions during the early stages of an experiment can significantly positively influence participant attitudes [8]. When designing study scenarios, researchers should be cautious of a possible primacy effect and similarly avoid atypical interactions at the beginning of the experiment if they anticipate it having a significant impact on participants' perceptions of the robot.

Minimizing Memory Biases: In order to counteract ordering bias, even in scenarios where abnormal interactions at the beginning or end of an experiment are needed, researchers may consider obtaining periodic subjective measurements throughout the interaction and averaging them. Our results suggest taking measurements after sub-interactions within a larger experimental interaction will provide higher measurement accuracy. This approach, however, can lead to counterproductive effects of participant boredom or fatigue if repeated survey instruments are administered too frequently. Adopting a "wash-out period" (increasing the time between participants' observations of major events) can also minimize memory bias [22]. However, this method necessitates lengthier experimental designs.

B. Practical Applications for HRI

The serial-position effect can be incorporated into risk-aware robotic planners in order to modulate user perception. Such planners could take the potential negative perception of risky actions into account while optimizing task performance. By adding a biasing factor to the robot's cost function, which assigns higher costs to actions that have a higher probability of leading to potential failures if they occur at the beginning or end of an interaction, the robot could automatically schedule its tasks to mitigate any negative primacy or recency biases and better highlight its successes, in domains such as in collaborative assembly [32] or social navigation [21]. If a failure does occur, the robot can artificially extend its interaction with a human beyond the failure to minimize its impact. This process is not limited to deliberately improving impressions of a robot's competence - it can also deliberately degrade such impressions by inverting the biasing factors. This capability is potentially useful for another critical application within HRI: trust calibration.

A crucial aspect of human-robot collaboration is a robot's ability to establish, develop, and calibrate trust over extended periods of time [2], [17]. It is suboptimal when a human fails to use a robot's capabilities when it would be advantageous to do so due to under-trust, not only because the robot's benefits are not properly utilized, but also because accidents can occur through the accumulation of Type II errors [6], [9]. Similarly, over-trust, where a human inherently accepts robot recommendations and actions even when they are suboptimal, can have serious consequences through the accumulation of Type I errors [18], [20]. Misaligned trust in a robot's true competence can have a significantly negative impact on the effectiveness and safety of a human-robot interaction.

Lewis et al. showed that the factors correlating best with trust in automation are a system's reliability (error rate), and the consequences of system faults [18]. Our results also showed a correlation between perceived robot competence and trust. The serial-position effect can therefore be leveraged to nudge a human's trust in a robotic system in the right direction when it is miscalibrated. For example, if it becomes clear that a human is overly relying on a robot with a high degree of task uncertainty, the robot's planner can intentionally showcase the robot's likely failures in a more prominent ordering. This can help the human move into a more deliberate thinking pattern where they break overreliance, taking appropriate action when necessary.

C. Conclusion

In this work, we investigated how the ordering of a robot's task failures and successes influences recalled perceptions of robot competence and trustworthiness via the serial-position effect. We ran an online human-subjects study (n = 53), where participants were shown video of a robot performing the same six tasks with the same levels of competence, with the ordering of robot successes, partial successes, and failures differing by experimental condition. Our results indicate that participants' post-experimental, recalled ratings of robot competence and trustworthiness were significantly lower when the robot failed at the end of the experiment, compared to when it failed at the beginning or the middle, indicating the presence of a recency bias. We also found that, despite this difference in recalled competence, the average of periodic subjective competence measures taken throughout the experiment remained stable across conditions. We expanded upon these findings to discuss potential implications for both designing experiments to minimize the impact of ordering effects, as well as designing robot planners that can utilize ordering effects to influence human perceptions of robot competence for the purposes of trust calibration.

REFERENCES

[1] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 2017, pp. 254–262.

[2] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[3] J. Deese and R. A. Kaufman, "Serial effects in recall of unorganized and sequentially organized verbal material." *Journal of experimental psychology*, vol. 54, no. 3, p. 180, 1957.

[4] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction*.

[5] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of changing reliability on trust of robot systems," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 73–80.

[6] S. R. Dixon and C. D. Wickens, "Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload," *Human factors*, 2006.

[7] B. L. Fredrickson and D. Kahneman, "Duration neglect in retrospective evaluations of affective episodes." *Journal of personality and social psychology*, vol. 65, no. 1, p. 45, 1993.

[8] I. Garnefeld and L. Steinhoff, "Primacy versus recency effects in extended service encounters," *Journal of Service Management*, 2013.

[9] M. Gombolay, X. J. Yang, B. Hayes, N. Seo, Z. Liu, S. Wadhwania, T. Yu, N. Shah, T. Golen, and J. Shah, "Robotic assistance in the coordination of patient care," *The International Journal of Robotics Research*, vol. 37, no. 10, pp. 1300–1316, 2018.

[10] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, 2019.

[11] G. Hoffman and X. Zhao, "A primer for conducting experiments in human–robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, 2020.

[12] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[13] D. Kahneman, "Evaluation by moments: Past and future," *Choices, values, and frames*, pp. 693–708, 2000.

[14] D. Kahneman, S. Frederick, *et al.*, "Representativeness revisited: Attribute substitution in intuitive judgment," *Heuristics and biases: The psychology of intuitive judgment*, vol. 49, no. 49-81, p. 74, 2002.

[15] D. Kahneman, S. P. Slovic, P. Slovic, and A. Tversky, *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[16] S. M. Kassin, M. E. Reddy, and W. F. Tulloch, "Juror interpretations of ambiguous evidence: The need for cognition, presentation order, and persuasion," *Law and Human Behavior*, vol. 14, pp. 43–55, 1990.

[17] B. C. Kok and H. Soh, "Trust in robots: Challenges and opportunities," *Current Robotics Reports*, vol. 1, pp. 297–309, 2020.

[18] M. Lewis, K. Sycara, and P. Walker, "The role of trust in human-robot interaction," *Foundations of trusted autonomy*, pp. 135–159, 2018.

[19] D. Lockton, "Cognitive biases, heuristics and decision-making in design for behaviour change," *Heuristics and Decision-Making in Design for Behaviour Change (August 5, 2012)*, 2012.

[20] M. B. Luebbers, A. Tabrez, K. Ruvane, and B. Hayes, "Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[21] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[22] S. Mukhopadhyay, M. D. Feldman, E. Abels, R. Ashfaq, S. Beltaifa, N. G. Cacciabeve, H. P. Cathro, L. Cheng, K. Cooper, G. E. Dickey, *et al.*, "Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study)," *The American journal of surgical pathology*, vol. 42, no. 1, p. 39, 2018.

[23] B. B. Murdock Jr, "The serial position effect of free recall." *Journal of experimental psychology*, vol. 64, no. 5, p. 482, 1962.

[24] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 33–42.

[25] T. Obo, C. Kasuya, S. Sun, and N. Kubota, "Human-robot interaction based on cognitive bias to increase motivation for daily exercise," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 2945–2950.

[26] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-hri"," in *Robust intelligence and trust in autonomous systems*. Springer, 2016, pp. 191–218.

[27] H. A. Simon, "Rational choice and the structure of the environment." *Psychological review*, vol. 63, no. 2, p. 129, 1956.

[28] M. J. Smith, I. Greenlees, and A. Manley, "Influence of order effects and mode of judgement on assessments of ability in sport," *Journal of sports sciences*, vol. 27, no. 7, pp. 745–752, 2009.

[29] R. H. Thaler and C. R. Sunstein, *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[30] T. Trainer, J. R. Taylor, and C. J. Stanton, "Choosing the best robot for the job: Affinity bias in human-robot interaction," in *Social Robotics: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings 12*. Springer, 2020.

[31] W. Tryon, "Chapter 11-clinical implications of network principles 3-12," *Cognitive Neuroscience and Psychotherapy*, pp. 501–561, 2014.

[32] Y.-S. Tung, K. Bishop, B. Hayes, and A. Roncone, "Bilevel optimization for just-in-time robotic kitting and delivery via adaptive task segmentation and scheduling," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication*.

[33] G. Verhulsdonck and N. Shalamova, "Creating content that influences people: Considering user experience and behavioral design in technical communication," *Journal of Technical Writing and Communication*, 2020.

[34] J. Xu and A. Howard, "The impact of first impressions on human-robot trust during problem-solving scenarios," in *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2018, pp. 435–441.