Affinity Diagramming with a Robot

MATTHEW V. LAW, Denison University, Granville, USA
NNAMDI NWAGWU, AMRITANSH KWATRA, SEO-YOUNG LEE,
DANIEL M. DIANGELIS, NAIFANG YU, GONZALO GONZALEZ-PUMARIEGA, AMIT RAJESH, and GUY HOFFMAN, Cornell University, Ithaca, USA

We investigate what it might look like for a robot to work with a human on a need-finding design task using an affinity diagram. While some recent projects have examined how human–robot teams might explore solutions to design problems, human–robot collaboration in the sensemaking aspects of the design process has not been studied. Designers use affinity diagrams to make sense of unstructured information by clustering paper notes on a work surface. To explore human–robot collaboration on a sensemaking design activity, we developed HIRO, an autonomous robot that constructs affinity diagrams with humans. In a within-user study, 56 participants affinity-diagrammed themes to characterize needs in quotes taken from real-world user data, once alone and once with HIRO. Users spent more time on the task with HIRO than alone, without strong evidence for corresponding effects on cognitive load. In addition, a majority of participants said they preferred to work with HIRO. From post-interaction interviews, we identified eight themes leading to four guidelines for robots that collaborate with humans on sensemaking design tasks: (1) account for the robot's speed, (2) pursue mutual understanding rather than just correctness, (3) identify opportunities for constructive disagreements, and (4) use other modes of communication in addition to physical materials.

CCS Concepts: • **Computer systems organization** → *Robotics;* • **Human-centered computing** → *Collaborative interaction;* • **General and reference** → *Design*;

Additional Key Words and Phrases: human–robot collaboration, affinity diagramming, design cognition, human–robot collaborative design

ACM Reference Format:
Matthew V. Law, Nnamdi Nwagwu, Amritansh Kwatra, Seo-young Lee, Daniel M. DiAngelis, Naifang Yu, Gonzalo Gonzalez-Pumariega, Amit Rajesh, and Guy Hoffman. 2024. Affinity Diagramming with a Robot. *ACM Trans. Hum.-Robot Interact.* 13, 1, Article 10 (March 2024), 41 pages https://doi.org/10.1145/3641514

1 INTRODUCTION

In this work, we ask what it might look like for a robot to assist a human designer in the process of affinity diagramming (Figure 1). Affinity diagramming is a common design method used to identify themes in unstructured data, such as finding user needs from interview transcripts. Data points,

This research was conducted with support from the Cornell University Center for Advanced Computing and has been supported by the Civil, Mechanical and Manufacturing Innovation Program of the National Science Foundation under NSF Award Nos. 1635253 and 1907542.

Authors' addresses: M. V. Law, Denison University, Granville, OH, 43023 USA; e-mail: lawm@denison.edu; N. Nwagwu, A. Kwatra, S.-Y. Lee, D. M. DiAngelis, N. Yu, G. Gonzalez-Pumariega, A. Rajesh, and G. Hoffman, Cornell University, Ithaca, NY, 14853 USA; e-mails: nan46@cornell.edu, ak2244@cornell.edu, sl2668@cornell.edu, dmd354@cornell.edu, ny46@cornell.edu, gg387@cornell.edu, ar883@cornell.edu, hoffman@cornell.edu.

ACM 2573-9522/2024/03-ART10
https://doi.org/10.1145/3641514

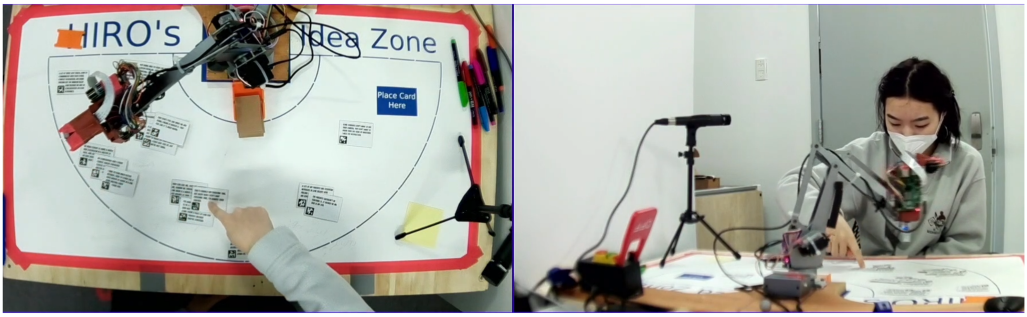


Fig. 1. HIRO, a robotic arm, builds an affinity diagram to organize user data with a human participant.

such as interview quotes, are transcribed onto paper notes that are then arranged, bottom-up, into spatial clusters combining related notes. The clusters are then labeled and used to organize insights about the data.

Recent studies [41, 50, 54] have explored the promise of humans and robots working together on various design tasks, e.g, through sketching or a shared tangible interface. These studies have uncovered complex design collaboration dynamics between the human and the machine. To the best of our knowledge, however, there has not yet been a study of how a physical robotic assistant might support the sensemaking aspects of the design process, e.g., need-finding and problem framing. Several digital systems have been designed to support affinity diagramming, as the method lends itself well to data mining and insights from natural language processing tools. However, studies have found that some designers still prefer to work with their hands and paper media (e.g., [9, 28]). A physical robot co-manipulating paper notes with the human designer could thus be well suited to the embodied nature of this design activity. Such a robot might manipulate shared design representations and use spatial reasoning to negotiate creative decisions with the human designer.

In this article, we explore this promise and set out to answer the following research questions:

- RQ1: How might working with a robot influence how humans construct affinity diagrams?
- RQ2: In what ways might affinity diagrams support creative collaboration between a human and a robot?
- RQ3: How should we design robots to support cognitive aspects of designing such as thematic clustering?

The contributions of this article are as follows:

- A robotic system that uses a pretrained language model and spatial information to construct affinity diagrams with a human
- Findings from a within-user study with novices at affinity diagramming ($n = 56$) suggesting that participants spent more time on the task when working with the robot without evidence of differences in cognitive load
- A set of eight themes describing participant experiences based on post-study interviews
- Design implications for future work exploring how robotic systems might support cognitive aspects of designing

2 MOTIVATION: WHY USE A ROBOT FOR AFFINITY DIAGRAMMING IN DESIGN?

Certain elements of designing are well suited for computational systems, especially in the solution-evaluation stage of the design process. For example, computers can quickly evaluate proposed designs at scale or search large solution spaces for optimal designs more efficiently than humans can. At the same time, the contextual and often amorphous process of formalizing design problems, happening earlier in the process, can make it difficult to access the benefits of algorithms. Unlike computer programs, human designers have the capacity to reason abductively about partial information and navigate ill-defined tasks [17], which are prevalent in the early phases of design.

Researchers have explored different ways to balance the benefits of computation with those of human intuition on ill-defined design tasks. For example, interactive genetic algorithms use human evaluators to guide a search algorithm in domains such as floor planning [59] or fashion design [43]. Other work frames computation as a tool to augment the curation of inspirational material [47], identify accessibility issues in user reviews [2], or support smart documentation of the design process [8].

The roles that computation should play in designing have evolved in parallel with theories about how designers think. For example, cognitivists such as Herbert Simon thought that design problems should be formalized and solved by search algorithms [63]. Others, such as Donald Schön, understood designing as situationally embedded, relegating computation to more targeted, supportive roles [61]. Recent proponents of enactive creativity (e.g., [14]) emphasize the role of interactions between agents in shared environments as the driver for creative work. In an enactive vision of human–computer creativity, humans and computers play off one another to improvise new ideas and perspectives.

Framing *interaction* as the engine of human–computer creativity has two significant implications for how we think about computers supporting design work. Firstly, it frames humans and computers as collaborative peers in the creative process rather than as users and tools. Secondly, it centers the interface through which humans and computer agents interact as a primary consideration in how to design such collaborations. Our interest in robots relates to each of these assertions, the former through the social interaction affordances of robot design, and the latter via their unique ability, as computational actors, to actively interface physically with humans and their thought processes.

2.1 Tangibile Interfaces to Support Material Human–Machine Design Interaction

There is strong theoretical precedent for centering social and, in particular, physical interaction in unstructured activities such as abductive design exploration. Situated theories of design cognition frame the designer's thought process as a journey through acting on the world, perceiving how it changes, and synthesizing new ideas. Schön calls this process a "reflective conversation with the materials of a design situation" [61]. Ingold argues that creative synthesis is a closed loop with active manipulation that follows the course of materials, just as each cut with a saw responds to the path of the previous cut through the wood. The tangibility of this process is paramount: "for there to be a rhythm," he writes, "movement must be felt" [38]. In design practices such as those in architecture, this active perception loop is evident through the pervasiveness of materially interactive practices such as sketching, which Goldschmidt describes as a sort of dialectic between different forms of perceiving the output of the pen [23].

Physically interactive and tangible interfaces are ideal for this sort of "thinking in action," as Tversky terms it [68]. In such interfaces, as Ishii observes, physical materials serve as coincident input and output spaces, cutting out a layer of mapping across spatial and modal discontinuities that might occur between a mouse and a screen [39]. Klemmer suggests that there may be value in interfaces that are closer to physical reality: "designing interactions that are the real world instead of ones that simulate or replicate it hedges against simulacra that have neglected an important practice" [45]. More concretely, storing the internal state of the system in physical materials, tangible interfaces afford what Ishii terms a *double feedback loop*, conveying immediate physical feedback and persistence beyond the digital domain [39]. This can be especially important for complex and unstructured work: "when thought overwhelms the mind," writes Tversky, "we put it into the world" [68]. Physical, tangible interfaces directly support the ability to do this.

In addition to their cognitive compatibility with unstructured and creative synthesis, tangible interfaces have certain implications for collaboration in human teams that may also apply to robots. Hornecker and Buur theorize that tangible materials can democratize collaborations by affording multiple similar access points to shared work [35]. Similarly, Ishii refers to the ability of tangible interfaces to multiplex space in ways that may resolve awkward concurrent interactions in graphical interfaces [39].

If these collaborative affordances extend to robotic agents, it opens the possibility of them physically manipulating material inputs and representations to act on a shared problem in ways that evoke a human collaborator. Robots are also well studied as social actors (e.g., [52]), laying the groundwork to position them as collaborative peers. Building on this premise, recent work has explored using robots that directly interface with designers through a shared physical environment to explore potential solutions to a design problem. In the following subsections, we discuss human–robot collaborative design systems in more depth and describe existing non-robotic efforts to support affinity diagramming with computation.

2.2 The Promise of human–robot Collaborative Design

Initial studies of **human–robot collaborative design (HRCD)** tools have suggested ways that human–robot interaction might benefit human–AI collaboration when designing solutions to a problem. Kahn et al. noticed an increase in creative expressions when a social robot offered humans creative prompts as they worked on a zen garden [41]. Law et al. identified social and creative collaboration challenges that occurred when a human and a robotic arm shared the same tangible design workspace [50]. They also observed instances in which participants interpreted the robot's physical behavior in social ways that influenced team dynamics. Building on this work, Lin et al. found that a co-sketching robot can increase satisfaction and inspire unexpected directions during ideation [54] when compared with a graphical co-sketching interface.

The still-open promise of HRCD lies in robots' potential, as physical actors, to communicate about a design situation through physical materials with human designers, using established norms about gestures and spatial reasoning. While the described studies have mostly focused on humans and robots working together to design solutions, designing is as much about problem framing as solution finding [17]. In this article, we explore what physical interaction with a robot and shared design materials, i.e. an affinity diagram, might look like within the problem-framing aspects of a human's design process, suggesting directions to advance HRCD.

2.3 The Practice of Affinity Diagramming and Technologies to Support It

Affinity diagramming is, loosely, a method of inductively clustering unstructured data into summary themes. It is historically associated with Kawakita Jiro, who developed the *KJ method* to construct hypotheses about raw data. The KJ method follows four primary steps [62]:

(1) Generate discrete labels representing pieces of data, e.g., transcribe user notes from interview quotes onto note cards.
(2) Group the labels, using a bottom-up process. For example, shuffle the user notes and successively add them to the diagram, building up clusters as themes emerge.

(3) Annotate the groupings and relationships between groups. This can be hierarchical, with subgroupings and relationships.
(4) Verbally explain the diagram and the insights that the team has culled from it.

In practice, professionals construct affinity diagrams for a variety of reasons, including eliciting diverse input (e.g., when brainstorming), organizing material, and analyzing data, and they frequently deviate from textbook procedures such as the KJ method defined above [28].

Despite the ability to construct affinity diagrams on-screen, some designers prefer using physical media for this activity. For example, Borlinghaus and Huber's design teams reported that they were uncomfortable working with pre-clustered notes, as they "were no longer 'thinking with their hands'" [9]. This has led some to suggest that technology that supports situated design activities such as affinity diagramming should augment tangible media and interaction, rather than supplanting it [27, 40].

With this in mind, affinity diagramming is a promising context for the study of conceptual design collaboration between a human and a robot. Much like sketching, affinity diagramming provides a way for designers to utilize space, gesture, and materials to think in and through the world around them. However, unlike many sketches [66], affinity diagrams are easy to interpret on their own. Textual notes and spatial relationships can be formalized straightforwardly, affording feasible common ground between a human and a machine. Nonetheless, the bottom-up nature of affinity diagramming creates flexibility around the meaning of notes and groupings. Designers allow themes and structure to emerge through the process rather than imposing any organization on the data. In short, affinity diagramming, as an open-ended creative conversation conducted through physical interactions with textual notes, is a convenient design activity through which to observe a human and a robot thinking together.

2.3.1 Technological Tools for Affinity Diagramming. Pertinent to affinity diagramming, several forms of interactive clustering have been explored, either incorporating human feedback into a clustering algorithm [4, 11, 19, 72] or vice versa [7, 18]. However, the efficacy of bridging human and computer approaches in the context of affinity diagramming is not always clear. Borlinghaus and Huber, for example, found that asking student design teams to work on preclustered data impeded students' willingness to dig into the meaning of the text [9].

A number of systems have been proposed to specifically support affinity diagramming in various applications, either fully digitally using a desktop or tablet interface or by augmenting paper media. *Designer's Outpost*, by Klemmer et al. [46], was an early attempt to improve inefficiencies within physical practice via hybrid tangible–digital interfaces supporting interaction using real Post-It notes. Among the insights the authors extracted through this system was that designers found proximity-based auto-grouping of notes superfluous since they could already see the note locations [44]. *AffinityFinder* was a tablet interface developed to help designers find notes in a large diagram. The user hovered the tablet over a diagram to perform keyword search [29]. Similarly, *AffinityLens* allowed participants to view rich informational overlays about notes and the relationships between them when hovering a phone over parts of the diagram [65]. While these two systems both preserve interactions with physical media, they impose a screen between the user and the diagram, reducing how directly users manipulate tangible information. *AffinityTable* is a tangible, rear-projection tabletop interface that allows designers to add notes using physical media; move, highlight, and zoom in on different portions of the diagram using physical tokens; and search notes based on authorship. The table also performs automatic spatial clustering and alignment, as well as image retrieval to augment the content of notes [22]. All of these systems modify the physical environment or provide digital overlays on the physical environment to augment the experience of affinity diagramming. To our knowledge, there is not yet a robotic system that directly builds affinity diagrams with a human.

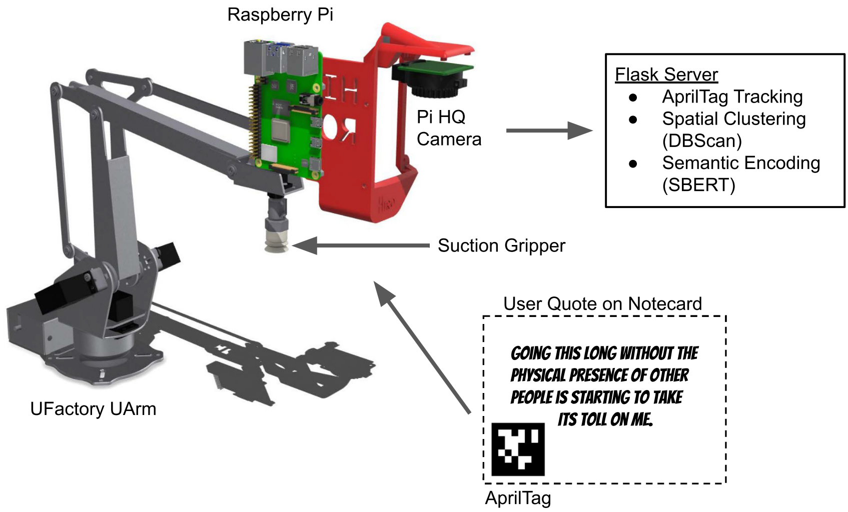

Raspberry Pi
Pi HQ Camera
Flask Server
● AprilTag Tracking
● Spatial Clustering (DBScan)
● Semantic Encoding (SBERT)
Suction Gripper
User Quote on Notecard
GOING THIS LONG WITHOUT THE PHYSICAL PRESENCE OF OTHER PEOPLE IS STARTING TO TAKE ITS TOLL ON ME.
UFactory UArm
AprilTag

Fig. 2. HIRO combines a UFactory UArm palletizer robot with a Raspberry Pi 4 and a Pi HQ Camera with a fish-eye lens. It uses the downward-facing HQ Camera to track April-tagged note cards on the surface in front of it. Note cards are clustered using their position on the surface via DBScan and semantic distance via a Sentence-BERT model on the backend. Once HIRO determines a new location for a card, it uses the UArm's suction cup to move the card to that location.

3 HIRO, A HOME IDEATION ROBOT

To study how a robot might engage with humans in conceptual design tasks such as affinity diagramming, we developed **HIRO (Home Ideation RObot)**, a tabletop robotic arm that collaborates on note card–based creative tasks with a human. For example, a person might sit across from HIRO and work through a stack of qualitative notes together with HIRO, taking turns adding them to a shared surface to gradually construct the thematic clusters of an affinity diagram. In the study described in this article, we gave participants a stack of note cards inscribed with user data to place in an affinity diagram. Rather than placing every card in the diagram themselves, participants could ask HIRO to select where a note should go in the diagram by placing the card in a designated zone. HIRO would then pick up the card, using a suction gripper, and place it in an existing cluster, accounting for both the way the cards are arranged in the diagram and the text inscribed on each card.

In the following, we map out HIRO's system architecture (Figure 2) and describe HIRO's functional and social behaviors with the system components that support each in the context of affinity diagramming.

3.1 System Architecture

HIRO is built on the UFactory UArm [69], which we use to move cards in the shared workspace. The UArm is a 4-DoF miniature palletizer arm with a suction gripper well suited for pick-and-place tasks. It is desktop scale, with a workspace of radius 36.5 cm. We instrumented the UArm's end-effector with a downward-facing Raspberry Pi "High-Quality (HQ)" camera, which contains a 12.3-megapixel Sony IMX477 image sensor. We mounted a fish-eye lens onto the camera to capture high-resolution, wide-angle images of the entire workspace. These images are used to track cards in the workspace and make decisions about where to place new cards in the diagram.

A script running on a Raspberry Pi 4 takes pictures of the workspace and moves the robot, offloading image and text processing to an external web server for performance reasons. The web server is built using the Flask web server framework [58]. The server localizes and associates text with all the note cards captured in an image using AprilTags [21, 73] printed on each card. AprilTags offer pixel-level position and orientation information about each note card, which we transform into workspace coordinates using the robot's pose. Each AprilTag is associated with a predefined quote via a lookup table. We compare relationships between quotes in the affinity diagram using a pretrained **Sentence-BERT (SBERT)** [60] model. SBERT is a variant of BERT [15], a language model previously used in the **human–robot interaction (HRI)** domain [37] that encodes semantic relationships in unstructured text using vectors. Based on the clusters of notes it identifies and encodes in the workspace, the server tells the Pi where to place new cards, using the method we describe below.

3.2 Affinity Diagramming with HIRO

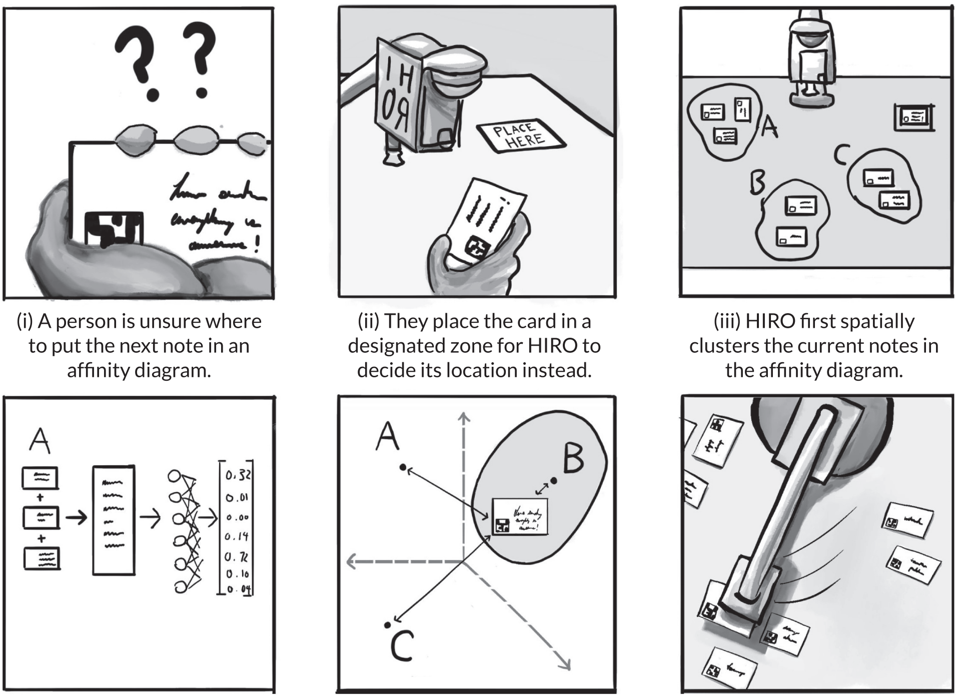
HIRO takes turns with a human to place note cards on a shared affinity diagram based on the diagram's current state. These turns are always initiated by the human; HIRO is not designed to proactively modify the diagram. Indeed, HIRO does not rearrange any cards that are already placed in the diagram, only adding cards on demand that the human places in a designated location in the workspace. HIRO also does not create new clusters, only adding cards to existing clusters as defined by the current arrangement of cards. These design choices were intended to simplify interactions between HIRO and the human for the purpose of this study. The process by which HIRO adds cards to the diagram is described below and illustrated in Figure 3.

3.2.1 Adding Cards to the Affinity Diagram. HIRO periodically takes photos of the workspace to locate any currently visible cards on the table. The designated "add zone" is a rectangular region marked to HIRO's left and the participant's right. Whenever HIRO locates a card in this designated zone, it samples a place for it to go in the diagram using both spatial and semantic information from the affinity diagram.

To accomplish this, HIRO first generates a map of current card locations from all detected AprilTags and uses DBSCAN [20], a density-based clustering algorithm, to group them into spatial clusters according to how close they are to one another in the workspace. We chose to use a density-based clusterer for flexibility in the number of spatial clusters HIRO could identify.

HIRO then converts each spatial cluster into an embedding representation as follows: it first looks up the text on each note in the cluster via a dictionary keyed on its AprilTag identifier; it then concatenates these texts into a single string document to represent that cluster. The resulting cluster "paragraph" is then encoded into a cluster embedding using the pretrained SBERT model 'paraphrase-mpnet-base-v2' [60]. HIRO finally encodes the text on the *new* card using the same SBERT model and assigns it to the closest cluster in the embedding space based on its Euclidean distance to the receiving cluster. Once it has selected a cluster, HIRO generates physical coordinates in the diagram to place the card by sampling a point near the center of all the cards in that cluster. With a target location, HIRO finally moves to add the card to the diagram. It first performs a gaze animation described below before turning to the add zone, moving down to pick up the card, then rotating to drop the card at the generated coordinates. This results in an auditory, physical, and visual interaction, as it includes the high-pitched noise of the servo motors, the humming of the vacuum, and the thump of the suction cup dropping onto the card it is picking up.

3.2.2 Social Behavior. HIRO is also programmed to perform social behaviors that don't directly involve affinity diagramming. Robots have long been studied as social actors. Previous work in HRCD suggests that humans will interpret a collaborative design robot's behaviors in social ways, e.g., collegial, dismissive, or antagonistic [50], even if they have not been designed that way. We designed the following behaviors to convey intentionality in HIRO's actions and emphasize HIRO's attention to the shared diagram.



(i) A person is unsure where to put the next note in an affinity diagram.

(ii) They place the card in a designated zone for HIRO to decide its location instead.

(iii) HIRO first spatially clusters the current notes in the affinity diagram.

(iv) HIRO stitches the notes in each cluster together and encodes them using a language model.

(v) HIRO identifies which encoded cluster is closest to the new note in the embedding space.

(vi) Finally, HIRO picks up the new card and adds it to the selected cluster.

Fig. 3. When a person wants HIRO to place a card (i), the individual puts it in a designated zone (ii). HIRO sees the card and spatially clusters notes in the diagram using DBSCAN (iii). HIRO then stitches together and encodes the text in each cluster using a pretrained Sentence-BERT model (iv). HIRO identifies the cluster closest to the new note in the embedding space (v), then picks up the note and moves it to the selected cluster (vi).

- Breathing: Rather than sitting idle between turns, HIRO continuously moves in a slow, pre-programmed periodic motion. HRI researchers have long suggested adding periodic motions to enhance animacy of an idle robot [25, 34]. Cuijpers and Knops found that idle motions such as breathing, swaying, or random head movements increased human social responses to a robot [13]. HIRO's palletizer mechanism constrains its end-effector to be parallel to the plane of its base at all times. Exploiting this, we created an idle breathing motion through which HIRO pulls its end-effector up and back, then down and forward again, without changing its angle with respect to the diagram on the shared work surface. This motion maintains the camera's orientation on the diagram and was intended to evoke task engagement, even when HIRO is not actively interacting with the note cards.
- Following: With a small probability, HIRO will also turn slightly towards the side of the diagram where it infers activity while waiting for a turn based on images from its downward-facing camera. We implemented this using a three-direction image classifier. Even if not always accurate, this motion was intended to convey attention towards the human's focus and is an analogue to Cuijpers and Knops's idle head movements [13].
- Gaze: Finally, when the user gives HIRO a card, it first turns and hovers over each cluster that it detects in the current diagram, pausing briefly above each one before pulling back and initiating its motion to pick up the new card. Since the images that HIRO captures from

the camera in its end-effector contain the entire workspace, this movement is not necessary to locate and process how notes are arranged in the diagram. Instead, it is designed to express intentionality and gaze attention to the spatial structure of the diagram.

4 USER STUDY: AFFINITY DIAGRAMMING USER NEEDS WITH HIRO

We ran a within-user study (n = 56) to examine how the process and experience of affinity diagramming with HIRO compared with working alone. Based on what we observed in initial pilot sessions of users affinity-diagramming with HIRO, we thought that working with HIRO might change the pacing of the task and perceived cognitive load. More importantly, we hoped to learn from participants' experiences what it was like to collaborate with a sensemaking robot through a physical diagram and identify key design considerations in their experiences.

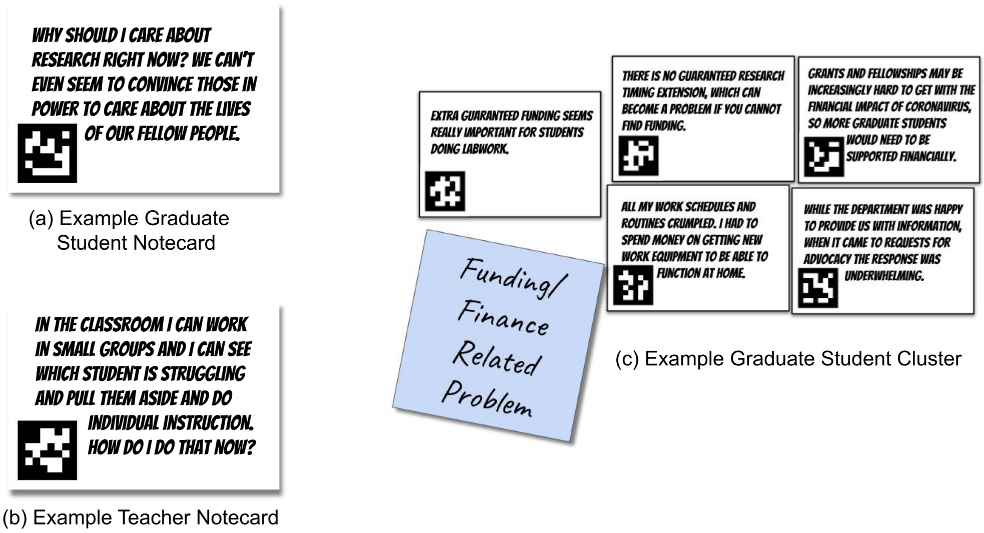
4.1 Study Design and Procedure

We used a mixed-methods approach to address our goals. To provide a simple quantification of the differences between treatments, we measured completion time and participants' self-reported cognitive load for each, evaluating hypotheses about the directionality of observed effects. These metric choices were informed by pilot studies. To get a richer understanding of how the experience of working with HIRO differed from working alone, we relied on participants' accounts via a semi-structured interview conducted at the end of each study session.

4.1.1 Hypotheses. We sought to quantify aspects of working with HIRO in order to establish baseline differences that distinguished the experience from working alone. Our hypotheses do not seek to establish the superiority of either condition, and the variables that we measured to evaluate them are appropriately broad for an initial study. For example, we measure the amount of time participants spend on the task using task completion time. Intuitively, the amount of time the participants spend could be influenced by the need to adjust to physically and cognitively working with a robot, including waiting for it to move, considering its point of view, and so on. Understanding these nuances would be necessary to determine whether an observed effect positively or negatively influenced the experience, and completion time is not granular enough to distinguish these nuances. Instead, as a first step, we sought to simply establish whether participants would spend longer to construct a diagram with the robot and flesh out potential explanations in our thematic analysis of participants' accounts.

Based on conversations and observations from a series of five informal pilot sessions in which we had users build affinity diagrams with HIRO, we sought to evaluate the following hypotheses:

(1) **H1:** *Participants will spend more time on the task when working with HIRO:* Taking turns with HIRO should enforce a slower pace, and participants may need to spend more time making sense of the data when negotiating with the robot than when working alone. As one pilot participant told us, "The robot was a lot slower, so it forced me to think more, because while I would be waiting for the robot to make its move, I would also be thinking, looking around more. But it was also a little frustrating."
(2) **H2:** *Participants will report higher mental demand when working with HIRO:* Participants may need to think more to account for and make sense of the robot's opinions and any differences with their own. As a pilot participant explained, "Working with the robot, sometimes it's challenging, because if the robot put a card in a cluster that I think doesn't belong to the cluster, it kind of disrupts my own definition of the cluster."
(3) **H3:** *Participants will report lower frustration when working with HIRO:* Participants may enjoy working with the robot or feel less pressured to do well on the task with its help. As a pilot participant put it, "I feel the robot made me a lot more relaxed because I felt I



WHY SHOULD I CARE ABOUT RESEARCH RIGHT NOW? WE CAN'T EVEN SEEM TO CONVINCE THOSE IN POWER TO CARE ABOUT THE LIVES OF OUR FELLOW PEOPLE.

(a) Example Graduate Student Notecard

IN THE CLASSROOM I CAN WORK IN SMALL GROUPS AND I CAN SEE WHICH STUDENT IS STRUGGLING AND PULL THEM ASIDE AND DO INDIVIDUAL INSTRUCTION. HOW DO I DO THAT NOW?

(b) Example Teacher Notecard

EXTRA GUARANTEED FUNDING SEEMS REALLY IMPORTANT FOR STUDENTS DOING LABWORK.

THERE IS NO GUARANTEED RESEARCH TIMING EXTENSION, WHICH CAN BECOME A PROBLEM IF YOU CANNOT FIND FUNDING.

GRANTS AND FELLOWSHIPS MAY BE INCREASINGLY HARD TO GET WITH THE FINANCIAL IMPACT OF CORONAVIRUS, SO MORE GRADUATE STUDENTS WOULD NEED TO BE SUPPORTED FINANCIALLY.

ALL MY WORK SCHEDULES AND ROUTINES CRUMPLED. I HAD TO SPEND MONEY ON GETTING NEW WORK EQUIPMENT TO BE ABLE TO FUNCTION AT HOME.

WHILE THE DEPARTMENT WAS HAPPY TO PROVIDE US WITH INFORMATION, WHEN IT CAME TO REQUESTS FOR ADVOCACY THE RESPONSE WAS UNDERWHELMING.

Funding/ Finance Related Problem

(c) Example Graduate Student Cluster

Fig. 4. Examples of note cards with quotes from (a) the graduate student dataset and (b) the teacher dataset, and (c) an example of a cluster that a participant created with HIRO for the graduate student data.

> had another companion who was also putting some thought into it. It sort of set the pace very well and it also here and there gives some good ideas."

In our full study, we measured H1 via completion time, and H2 and H3 via subscales of the **NASA Task Load Index (TLX)** [31]. We set a sample size of n = 56 to evaluate these hypotheses via a power analysis using pilot data. In order to support transparency and reproducibility, we preregistered our hypotheses and quantitative analysis plan (https://aspredicted.org/HN7_6RK). There were no additional dependent variables measured to evaluate our hypotheses.

4.1.2 Task and Treatments. Each study participant constructed two affinity diagrams: one on one's own and one with HIRO. Each affinity diagram was constructed using one of two different datasets: (1) a survey of graduate student experiences during the COVID-19 pandemic [51] and (2) an interview study with teachers about their experiences during the same pandemic [70]. Example quotes from each dataset can be found in Figure 4. We selected 28 quotes from each dataset and printed them on two stacks of cards. For each treatment, we asked participants to organize the quotes from one stack into clusters representing that community's primary needs. Figure 5 and Figure 6 illustrate the physical setup and procedure for each experiment, respectively.

The note cards for a particular treatment were stacked face down in a small container directly in front of HIRO's base. When working alone, participants placed every card in the diagram themselves, with HIRO sitting in front of them but shut off. When working with HIRO, participants took turns placing note cards with HIRO, which was introduced in the study script as follows: "For this set of user data, you will be working with HIRO, an ideation robot. Instead of adding every card to the affinity diagram yourself, you and HIRO will take turns placing the next card. When it is HIRO's turn to place a card, take the next card off the deck and put it in the blue rectangle on the top right of the workspace. HIRO will then add the card where it thinks it makes the most sense in the diagram." Participants were given discretion over which cards, when, and how many to share, and were reminded that they could move any card in the diagram at any time during the process.

In order to train participants on how to interact with HIRO without biasing how they built conceptual clusters, the researcher briefly demonstrated working with HIRO to sort a set of cards by

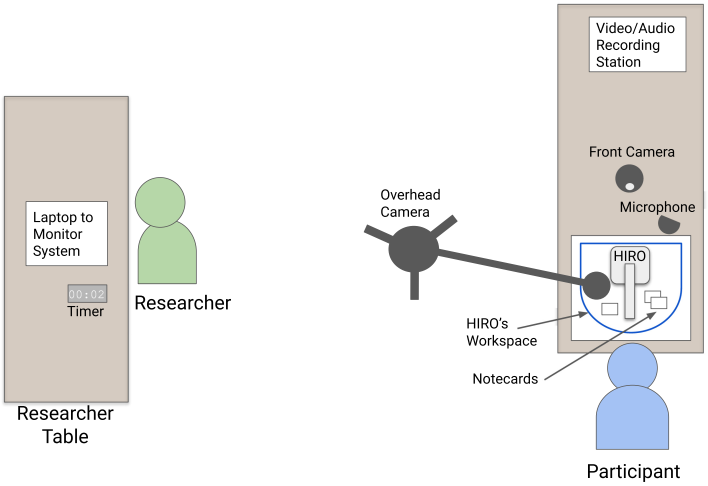


Fig. 5. Experimental Setup: Studies were run in a closed room with video and audio recording equipment. Participants sat on one side of the room at the end of a table facing HIRO. Diagrams were constructed on a placemat with an outline of HIRO's approximate workspace reach. During the study, the researcher sat across the room at another table to monitor the system.

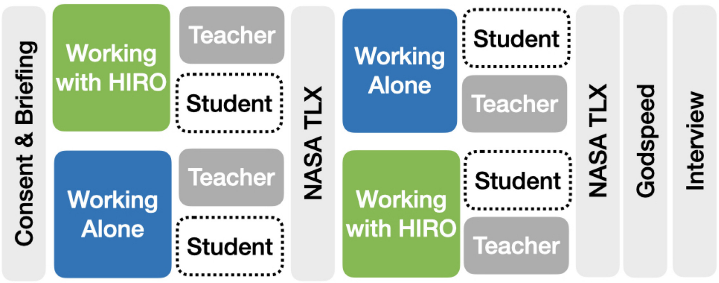


Fig. 6. Experimental Procedure: After consent and briefing, participants completed two affinity diagrams in turn, one with HIRO and one alone. Each diagram was built from different user data (teachers or graduate students). Treatment and dataset order were randomly assigned. After each treatment, participants completed a NASA TLX survey. At the end of the study, they completed a Godspeed survey and a semi-structured interview.

color. These cards did not have text on them and were clustered using RGB values. We randomized both treatment order and the assignment of dataset to treatment in order to counterbalance effects from either learning or differences between the datasets (see Figure 6).

4.1.3 Procedure. The full study procedure is shown in Figure 6. Studies were run in a closed, windowless room, with tables and video and audio recording equipment (Figure 5). One table was set up lengthwise on each side of the room. HIRO was placed, facing outwards, at the end of one table, on top of a placemat with outlines of its reachable workspace and the note card "add zone." Participants sat at the end of the table facing HIRO. Each study was recorded with two cameras: one on a tripod and boom that captured a top-down view from above the shared workspace and one on a short tripod facing the participant. A directional microphone was also pointed towards the participant, on the table next to HIRO. The other table was set up on the other side of the room–here, the researcher would sit with a laptop to start, stop, and monitor the robot. The researcher would answer any questions the participant had during the study.

Each study was run according to a detailed checklist and script enumerating the sequence of steps to set up the system and recording equipment, introduce the participant to the study and the robot, and initiate and conclude each treatment. Following the informed consent process and demographic survey, the researcher invited the participant to sit in front of the robot and briefly explained affinity diagramming as “a tool that designers use to organize information into themes and identify different needs in target communities.” They then introduced the user data and task as follows: “Today you will be working with quotes from surveys and interviews with two different demographic groups, teachers and Cornell grad students, about their experiences during the COVID-19 pandemic. Each quote is recorded on a note card. Your job is to go through these quotes one by one and organize them into clusters as you go. Before you finish, you will also label each cluster with a Post-It note describing the theme or need that cluster represents. The idea is that a designer can then design something to address the needs that you’ve identified for that group of people. You will do one affinity diagram on your own and one with a robot.”

Before each treatment, participants were reminded that their task was to cluster the note cards into themes. They were also reminded that they could have as many or few clusters as they wanted and could move any card around the diagram at any time. Finally, they were asked to read the cards and think aloud as they worked. An example cluster from a participant’s affinity diagram is shown in Figure 4(c).

Participants took as much time as they needed to complete each affinity diagram, timed by the researcher. If participants forgot to label their clusters, they were reminded to do so before completing the diagram. On completion, participants verbally explained their groupings to the researcher, then filled out the NASA TLX [30, 31] via a digital survey, measuring cognitive load over six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. At the conclusion of the study, participants completed a Godspeed survey [6] evaluating how they perceived HIRO as a social robot.

4.1.4 Semi-structured Interviews. Each study finished with a semi-structured interview comparing the experience of affinity diagramming with the robot and working alone. **Semi-structured interviewing (SSI)** is a frequently used method in qualitative interviews that falls between the closed structure of a survey and the complete open-endedness of unstructured methods. SSI emphasizes conversational interviewing around a list of topics (an “interview guide”) rather than following a fixed questionnaire [1]. Often, this involves probing followup questions or judiciously letting the interviewee drive the dialogue. Our researchers were trained in this approach and worked off the list of topics in Table 1. The interviewer would typically begin with the first question on the list, asking participants to broadly compare the two treatments. They would then move to cover the remaining topics in the order that the conversation dictated. Researchers were encouraged to probe for details or specific examples to flesh out participants’ responses, as suggested in the interview guide. There was no set length for the interviews, which lasted an average of 9 minutes and 44 seconds and typically concluded with the interviewer giving the interviewee an opportunity to ask questions about the study.

4.1.5 Participant Recruitment and Demographics. Participants were recruited using mailing lists, fliers, and a university credit system. The study was reviewed by the university **institutional review board (IRB)** and all participants gave written informed consent before participating. We obtained release signatures for any photos or videos that are used in publications.

Table 1. Topics Covered in the Post-study Semi-structured Interviews, Along with Opportunities for Follow-up Questions

Interview Topics
How was working with the robot versus alone? Can you illustrate with specific examples?
Can you describe your process when working with the robot? (If doesn't touch on personal process: How did that differentiate from working alone?)
Did anything surprise you when working with the robot? Can you give a specific example?
Would you prefer working with the robot or alone to build an affinity diagram? (If response is unclear: Do you think you performed better with or without the robot?)
To what extent do you feel HIRO understood what you were thinking? To what extent did you understand what HIRO was thinking?
What percentage of the final outcome would you attribute to the robot?
Did you ever disagree with the robot and, if so, how did you handle it? Why did you handle the disagreement this way, e.g., why did you choose to override the robot or not?
Do you think working with the robot was different than what you would imagine working with a human on this task would be? How so?

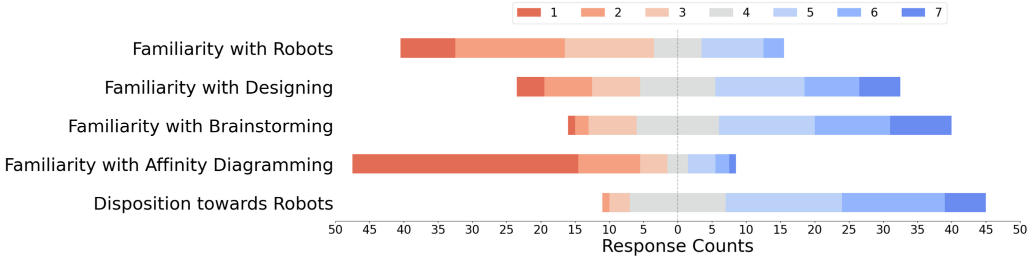


Fig. 7. Participants rated their familiarity with robots, designing, brainstorming, and affinity diagramming, as well as their feelings about robots, on a scale of 1 to 7. Participants reported low familiarity with affinity diagramming and were well disposed to robots.

Participants were between the ages of 18 and 62 years old ($\mu = 22.8$, $\sigma = 6.12$). A total of 43 participants identified as female, 11 as male, and 2 preferred not to respond. We asked participants to rate their familiarity with robots, designing, brainstorming, and affinity diagramming, as well as their feelings about robots, on a scale of 1 to 7. Their responses, plotted in Figure 7, indicate that participants were unfamiliar with robots and affinity diagramming but tended to be more familiar with designing and brainstorming in general. Also, participants were well disposed towards robots prior to the study.

5 FINDINGS

In the course of our study, we collected survey data on per-treatment cognitive load via NASA TLX and perceptions of HIRO via Godspeed, as well as task completion time, audio recordings of post-study interviews, video and audio data of the task, and the cluster labels that participants generated for each affinity diagram they made. Below, we present quantitative findings to evaluate each of our study hypotheses. We also present post-hoc analysis of how participants split the task of adding notes to the diagram with HIRO and immediately responded to its choices, the number of clusters that they created from the user data, and their perceptions of HIRO as a social robot. Finally, we discuss eight themes that we identified in our post-study interviews.

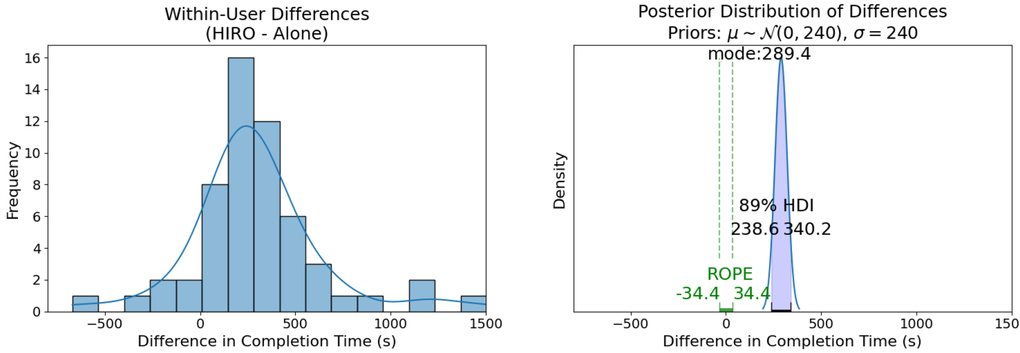


Fig. 8. Participants tended to spend more time on the task when working with HIRO than alone. The left plot shows the distribution of differences between treatments for each participant. On the right, the posterior distribution of these differences, with indicated priors, is centered at 289.4 seconds longer with an 89% HDI from 238.6 to 340.2 seconds (shaded interval), outside the ROPE of ±34.4 seconds (green dashed lines).

5.1 Experimental Findings

5.1.1 Task Completion Time. During each study, the researcher hand-timed each treatment, from the first card draw until participants declared that they were finished. Some times were manually corrected for technological issues or delays. Participants spent an average of 18 minutes and 26.3 seconds overall across treatments. They spent, on average, 294.6 more seconds (4 minutes and 54.6 seconds) working with HIRO than working alone, with a standard deviation of 343.9 seconds, although six participants spent less time working with HIRO than working alone.

We adopted a Bayesian approach to analyze this effect, modeling the mean difference in completion time using a normal likelihood function with a weakly informed normal prior over the mean centered at $M_\mu = 0$ with standard deviation, $S_\mu = 240$ seconds and a fixed standard deviation, $S_y = 240$ seconds. To evaluate an effect in completion time, we compared an 89% **highest density interval (HDI)** for the posterior distribution of the mean difference in completion time between treatments to a **region of practical equivalence (ROPE)** of 0.1 standard deviations around zero [49].

When using a normal likelihood function with priors of $\mu \sim \mathcal{N}(M_\mu, S_\mu)$ and fixed standard deviation S_y, the posterior distribution on μ is a normal distribution with closed-form solutions for the mean and standard deviation [48]. The posterior distribution over the mean difference in completion time, given our data and priors (Figure 8), is centered on 289.4 seconds, with an 89% HDI from 238.6 to 340.2 seconds, well outside the ROPE of −34.4 to 34.4 seconds. This suggests that, with 89% probability, the mean difference in completion time between treatments is between 238.6 seconds (3 minutes and 58.6 seconds) and 340.2 seconds (5 minutes and 40.2 seconds) longer when working with HIRO. As such, there is strong evidence to support **H1**. Prior and posterior predictive checks for this analysis can be found in Appendix A, as well as a sensitivity analysis over different parameterizations of the priors.

5.1.2 Cognitive Load. **H2** and **H3** predicted that participants would report higher mental demand and lower frustration, respectively, two subscales of NASA TLX, when working with HIRO. We again applied a Bayesian analysis to the differences in reported TLX scores between treatments, using a normal likelihood with a normal prior over the mean and fixed standard deviation. To model the effect in each of these two subscales, we parameterized the mean prior with mean zero and standard deviation 20, and fixed the standard deviation at 20.

As can be seen in Figure 9, again using a ROPE of 0.1 standard deviations around 0, our data does not support either hypothesis. The posterior distribution for mental demand is centered around a

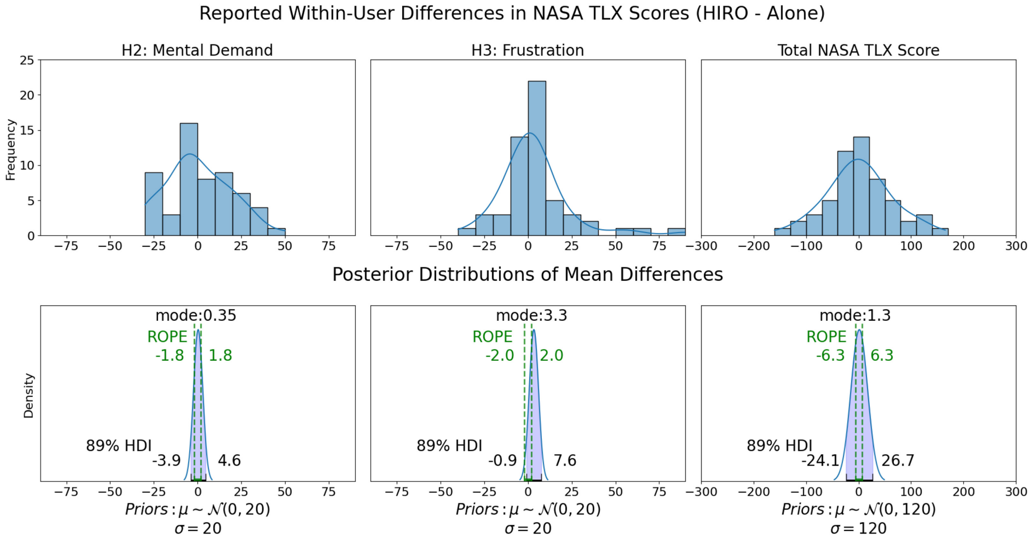


Fig. 9. Study data did not support a conclusive difference between participants' cognitive load along the two NASA TLX [30, 31] dimensions of mental demand and frustration when working with HIRO, as seen here in the distribution of observed differences (row one) and posterior over the mean difference with indicated priors (row two). Total TLX scores are shown for reference on the right. As the HDI (shaded interval) overlaps the ROPE (green dashed lines) in each posterior, there is insufficient evidence to conclude credible effects.

difference of 0.35 points on a 100-point scale. The frustration posterior is centered at 3.3 points. Both HDIs overlap the ROPE significantly, inhibiting any conclusions about the effect of HIRO on mental demand and frustration. Prior and posterior checks, as well as sensitivity analysis over the prior parameters, can be found in Appendix B. Beyond our hypotheses, we also did not find evidence to support meaningful effects across the remaining four subscales of TLX (see the right-hand column of Figure 9).

5.2 Dividing Card Placements with HIRO

To get a sense of how participants divided the task of constructing their affinity diagrams with HIRO and how they immediately reacted to HIRO's choice of placement, we analyzed card placements in each session. Specifically, we video-coded the number of cards that HIRO placed in each session (placements) and the number of times that participants chose to immediately change HIRO's placement (reversals). Placements were defined as time segments beginning with HIRO picking up a card and ending with the participant taking a new card. Reversals were defined as time segments starting with participants picking up a card that HIRO just placed and ending with them placing it elsewhere.

Two authors independently coded the study videos using ELAN [74]. This coding primarily relied on video data, using audio to resolve ambiguous cases. The coders met once to discuss edge cases they encountered. For instance, we classified instances in which HIRO physically failed to pick up the card and the participant, inferring what it meant to do, moved the card for HIRO, as placements. The coders met once more to check and fix discrepancies and identify disagreements. Twelve disagreements between the coders were then independently resolved by a third author, resulting in a final set of 797 placements and 241 reversals.

As seen in Figure 10, participants let HIRO place an average of 14.2 cards (σ = 4.77), which corresponds to roughly half the cards in the stack. Participants immediately modified an average

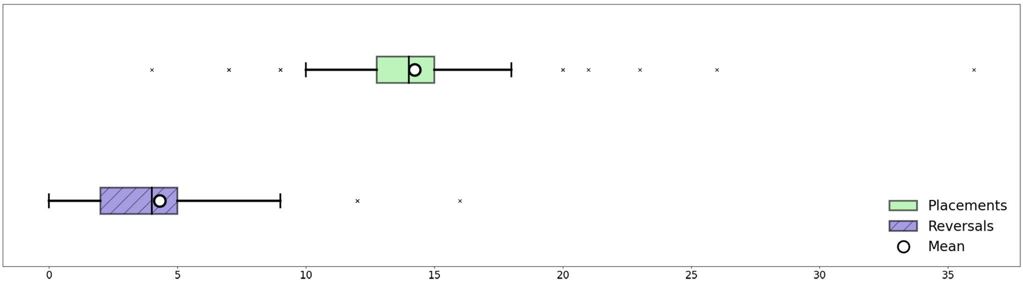


Fig. 10. HIRO made an average of 14.2 card placements per session, with participants immediately modifying an average of 4.3 placements. Participants were given control over how to distribute placements between themselves and HIRO. Some participants asked HIRO to place cards that were already in the diagram, which could lead to sessions in which the number of times HIRO placed a card exceeded the number of cards in the deck.

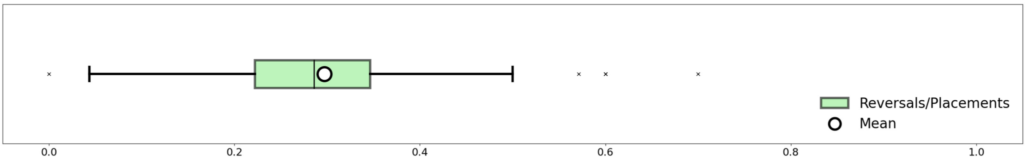


Fig. 11. On average, the ratio of HIRO's moves that participants immediately modified to its overall placements was 0.297, spanning from 0.0 to 0.7.

of 4.3 placements ($\sigma = 2.95$). Placement numbers can be higher than the total of 28 in the stack if, for example, participants took a card out of the current diagram and asked HIRO to re-place it somewhere. Overall, we found a great range in how participants divided and responded to card placements. HIRO placed as few as 4 and as many as 36 cards in a session; participants immediately modified as few as 0 and as many as 16 of these placements. For each participant, the ratio of reversals to placements averaged 0.297 ($\sigma = 0.143$), ranging from 0.0 for a session in which HIRO placed 13 cards to as high as 0.7 for a session in which it placed 10 cards (Figure 11).

5.3 Topics Identified by Participants

We counted the topics that participants covered in their groupings as a way of characterizing the diagrams that they produced. Across both datasets, participants tended to create slightly fewer clusters when working with the robot ($\mu = 5.61$, $\sigma = 1.87$) than when working alone ($\mu = 6.46$, $\sigma = 1.89$). Figure 12 plots the distributions of the number of clusters participants created in each treatment, overall and broken down by the user data they were affinity diagramming.

5.4 Perceptions of HIRO as a Social Robot

At the end of the study, participants completed a Godspeed [6] survey measuring their perceptions of HIRO as a social robot. Godspeed asks participants to rate their impressions of a robot on a scale of 1 to 5 between pairs of adjectives describing its anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Participants' responses are summarized in Figure 13.

Participants' perceptions of HIRO reflected its design as a non-humanoid cognitive robot. Overall, participants rated HIRO as more likeable ($\mu = 4.09$, $\sigma = 0.85$) and intelligent ($\mu = 3.78$, $\sigma = 0.84$) than anthropomorphic ($\mu = 2.73, \sigma = 1.11$) or animate ($\mu = 3.22$, $\sigma = 1.00$). Drilling deeper reveals some interesting contrasts within these subscales. For example, when thinking about

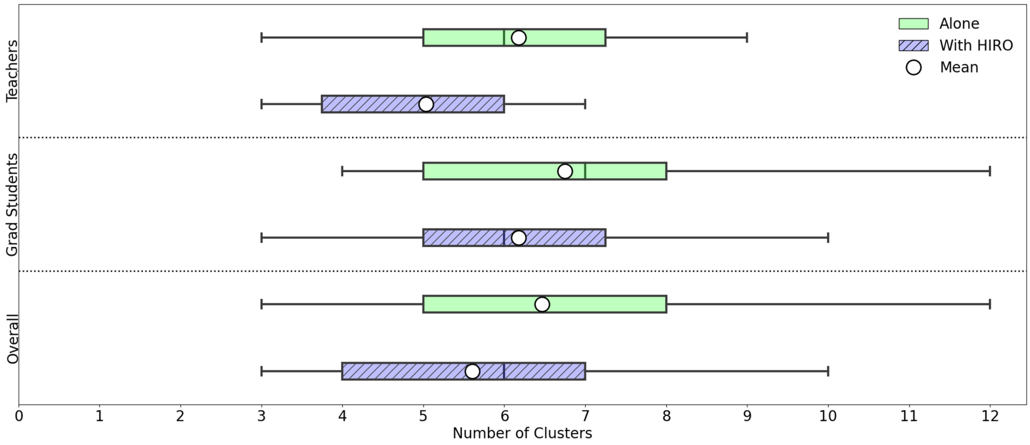


Fig. 12. Participants tended to create slightly fewer clusters when affinity diagramming with HIRO ($\mu = 5.61$, $\sigma = 1.87$) than when working alone ($\mu = 6.46$, $\sigma = 1.89$), as seen in the third row above. This trend was small but consistent across both sets of user data that participants worked with (rows one and two).

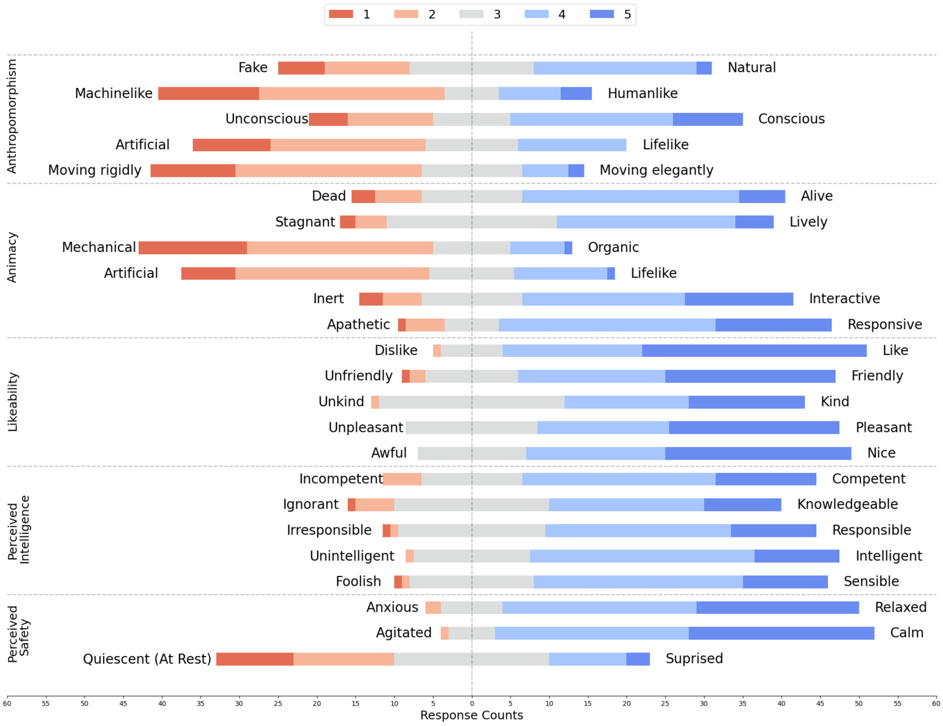


Fig. 13. Participants rated HIRO with higher likeability and perceived intelligence than anthropomorphism and animacy. Within the latter two categories, they rated HIRO as more natural and conscious than humanlike, lifelike, or elegant, and as more alive, lively, interactive, and responsive than organic and lifelike.

anthropomorphism, participants rated HIRO as machinelike, artificial, and moving rigidly, yet at the same time natural and conscious. Likewise, in terms of animacy, they rated HIRO as mechanical and artificial but also alive, lively, interactive, and responsive. The perceived safety subscale asks participants to rate their own emotional state rather than their impressions of the robot. Here, participants largely reported feeling relaxed and calm after working with HIRO but more quiescent than surprised.

6 THEMATIC ANALYSIS OF POST-STUDY INTERVIEWS

Each study concluded with a semi-structured interview comparing the experience of working with HIRO to working alone (see Section 4.1.4). In these interviews, 33 participants said they preferred to work with HIRO, 14 preferred to work alone, and 9 did not express a clear preference. We extracted a set of themes from our post-study interviews using a thematic analysis conducted via affinity diagramming. Precedents for this analysis method include, for example, Nunes et al. [56] and Lucero [55].

To perform this analysis, we first used a third-party service to transcribe the interview audio. Two authors then segmented the transcripts into notes bearing quotes representing the key ideas articulated by each participant. Three authors then collaboratively affinity diagrammed these user notes to identify the eight themes that are presented in Table 2 and described below. This affinity diagramming took place across 5 weeks with the three authors, who were not co-located, meeting in seven virtual collaborative sessions. The notes were divided randomly among the three authors, who each took turns adding them one at a time onto a shared board, where clusters represented thematically similar notes. Authors temporarily labeled clusters as the diagram developed, but would also reorganize if a more compelling grouping emerged as new notes were processed. Disagreements were handled as they arose by voting. At the end of this process, the affinity diagram contained labels at different levels of abstraction, which the authors organized into the eight themes discussed below. Themes are illustrated using representative interview quotes; the speaker for each quote is identified as $Px(yP, zR)$, where x is the participant identifier, y is the number of HIRO card placements in their session, and z is the number of reversals.

6.1 Second Perspectives Provided by the Robot

Participants identified several ways that having HIRO as a second perspective affected the experience of affinity diagramming. Some participants appreciated HIRO as an additional decision maker, especially when they were unsure about how to cluster certain notes. "I felt like I was just going along a straight one-way road when I'm working alone," said P1 (20P, 2R), "but with the robot, I felt like I had some kind of company with the thinking process and some optional solutions to rely on when I'm stuck." HIRO also led some participants to explore more alternatives than they did alone. P40 (14P, 2R): "I felt like I was changing a lot more when I was with the robot, [...w]hereas, when I was working [alone], I didn't switch a card once." Even when HIRO's perspective wasn't convincing, evaluating it reinforced their confidence in their own decisions. P52 (16P, R8), despite disagreeing frequently with HIRO's choices, said "I also feel more secure, because I actually thought more about it."

HIRO's second perspective helped participants resolve uncertainty. "Sometimes in my mind," described P13 (13P, 4R), "I'm not really sure which group this would go to. And then, the robot would decide, and I'd be like, oh, that make sense." P26 (14P, 4R): "I don't feel very confident working with open-ended questions by myself. Another person's decision would definitely help. And HIRO is pretty intelligent to me and I appreciated his help." This was especially important at the beginning, as P11 (14P, 3R) told us: "I was indecisive at first. [...] But with the robot [...] it dictated a path [...] it was much easier to categorize things in that way." For others, simply seeing

Table 2. Themes Extracted from Post-study Interviews, Aspects of Individual User Experiences that Describe Each Theme, Corresponding Design Guidelines (Denoted DG) and Representative Quotes

Theme	Participants felt that . . .	DG(s)	Representative Interview Quote(s)
Second Perspectives	• Having an additional decision-maker introduced a new perspective. • Seeing HIRO's choices could reduce uncertainty. • Checking or fixing HIRO's choices took extra effort. • It could be stressful to have your work checked.	DG2, DG3	• "With the robot, I felt like I had some kind of company with the thinking process and some optional solutions to rely on when I'm stuck." • "I felt like I was changing a lot more when I was with the robot, [. . . w]hereas, when I was working [alone], I didn't switch a card once." • "I had to think harder about it to figure out whether or not I wanted to change the way I initially approached it." • "I felt like when you sit at the kitchen table and do math homework with your dad. And you know that they're watching, knowing the correct answers, and thinking you're stupid for making the wrong decisions."
Mutual Understanding	• Shared understanding with HIRO was dynamic over the course of the task. • Disagreements could be reasonable or inexplicable. • Repeated or inexplicable disagreements could undermine collaboration.	DG2	• "We were learning a common language at that point. We were building off of things that can be used to inform understanding. But before that critical mass of information occurs, I did not understand what HIRO was doing."
Cognitive Load	• It took additional effort to include HIRO in the task. • Working with HIRO could reduce overthinking. • HIRO could help to automate the task.	DG3	• "Myself, I could make the organization make sense in my head, but [with HIRO] I wanted to make sure it was clear so they could detect the different clusters," • "Alone . . . there are different aspects I would rethink over and over again, compared to if the robot would just take the card and place it in a particular stack [and] I would try to think of the why and not." • "The robot just became more of a time-saving machine. It's, like, a way to double my bandwidth."
Pacing and Flow	• Waiting for HIRO was slow or the extra thinking required more time than working alone. • HIRO moved too fast and distrupted their desired pace. • The reduced pace allowed them to reflect or plan as they went.	DG1	• "You had to wait for the robot a lot. Sometimes, it would take a while to go and get the card and figure out where it wanted to put it." • "Definitely it takes more time to work with a robot, which is good because I think it allowed me to re-examine my decisions."
Enjoyment	• Working with HIRO was fun. • HIRO was socially engaging. • Figuring out HIRO was like playing a game.	—	• "It was friendlier, I don't know how else to put it, but it was more enjoyable." • "I feel like we connected. I feel like we're friends. I would be very sad if HIRO doesn't like me."
Nonverbal Communication	• Verbal communication could simplify collaboration mechanics. • Verbal communication could support more explanation, debate, and higher-level discussion of concepts. • The lack of mechanisms to argue with HIRO was beneficial.	DG4	• "I can't talk to the robot and explain my thinking. And if HIRO was a human, I would ask them what they were thinking." • "I'm a little introverted and he doesn't talk back but he gets the job done." • "It feels like my thoughts, my worries, my insecurities are just put out there and I just don't have anyone to reassure me or guide me in the right direction."
Perceived Intelligence	• HIRO's ability to contribute was bounded according to how they understood its reasoning to function internally. • HIRO's movements influenced how they perceived its intelligence and animacy.	DG2	• "It went about as I expected it to go, a robot sorting human concerns. People obviously innately understand human struggles more than many robots are capable of." • "[. . .] I could reason myself and get to the same place he did. So, there was more thought I guess, and less algorithm." • "It's very mechanical, but also I could see a human quality kind of like pondering as he went."
Roles and Power Dynamics	• HIRO could play different collaborative roles, from a reference tool to a decision-maker. • It could be more or less appropriate for HIRO to either define clusters or sort cards into clusters. • Differences in knowledge, physical constraints, or their empathy towards HIRO contributed to a sense of power dynamics between them and the robot.	DG1, DG3	• "I was the sole person making categories, then [HIRO is] somebody who's just helping me sort things." • "[. . .] instead of having to generate categories myself I could either agree or disagree with the robot." • "[. . .] I definitely saw our power dynamic; the robot had more power and knowledge in this area." • "I think my reasoning trumps his, but not in a disrespectful way [. . .]" • "It felt like I was offending him, because it can't defend itself, right?"

HIRO's opinion could help clarify their decisions. As P28 (14P, 4R) explained, "When HIRO picked [the note] up, [. . .] just the act of watching it make a decision, I was, like, okay, I actually know where I want this to go."

At the same time, many participants found it more stressful or less efficient to deal with HIRO's perspective on the data. Participants complained about having to double check HIRO's work, inhibiting their ability to work intuitively. Instead of going off "the first thing that popped into my head," said P45 (15P, 9R), "I had to think harder about it to figure out whether or not I wanted to change the way I initially approached it." This was particularly challenging when participants frequently disagreed with the robot. P51 (10P, 7R) told us that working with HIRO was a burden, "because I was trying to fix things I didn't agree with." Some participants found it more stressful to have HIRO checking on *their* work. P56 (14P, 4R) remembered feeling judged in HIRO's presence. As they described it, "HIRO is, like, a genius machine that has been programmed to know how to do this. And I'm just coming in as an undergraduate in college. And so, I felt like when you sit at the kitchen table and do math homework with your dad. And you know that they're watching, knowing the correct answers, and thinking you're stupid for making the wrong decisions."

6.2 Mutual Understanding: Getting on the Same Page with HIRO

Participants perceived varying degrees of mutual understanding with HIRO about the structure of the diagrams they were constructing together. Some felt on the same page as HIRO early on in the process, while others struggled to understand HIRO or felt like HIRO didn't understand their thought process.

This sense developed over the course of the task. "As the experiment went on," P14 (13P, 3R) recalled, "HIRO started placing cards in clusters that I started agreeing with more. So, in that sense, I felt like he was learning and paying attention." For P28 (14P, 4R), mutual understanding developed through collaboration: "We were learning a common language at that point. We were building off of things that can be used to inform understanding. But before that critical mass of information occurs, I did not understand what HIRO was doing." Some participants never reached that tipping point with HIRO, leaving them frustrated and confused. For example, P6 (18P, 8R), who was excited to work with HIRO, at some point decided: "I'm very confused and I'm trying to generalize the idea behind this category but I have been struggling; so I'm just going to leave HIRO out of this, I'm going to start deciding on my own."

Participants tended to categorize disagreements with HIRO as either reasonable or inexplicable. Reasonable disagreements, for example, could occur when participants were conflicted about where to place a card. "I moved it," P3 (14P, 1R) explained, "But I could see why HIRO put it there because I was also kinda debating between those two categories." When participants failed to parse HIRO's reasoning, it could undermine mutual understanding, even if they mostly agreed with its placements. "I feel like I didn't understand HIRO . . . we agreed on a good amount of items but there were some choices that I was very confused by," said P2 (15P, 6R).

While repeated disagreements could erode trust, some participants worked to repair gaps in mutual understanding with HIRO. For example, P1 (20P, 2R) described reversing an earlier decision to override HIRO as P1 started to see its point of view, saying, "I thought it was an error, and then it made more sense after I knew what the other cards were and realized the robot was probably right." When P23 (14, 4R) couldn't make sense of HIRO's choices, P23 tried to make clusters clearer, remarking that, "HIRO was pretty much telling me this category doesn't make sense."

6.3 How HIRO Affected the Cognitive Load Required for the Task

Despite the minimal differences measured by NASA TLX, participants described increased or decreased aspects of cognitive load when working with HIRO. As discussed in Section 6.1, it could be burdensome to check HIRO's work or more stressful to think with HIRO watching. Participants also devoted effort to ensuring that HIRO could see and understand how they were organizing the cards. This manifested at a physical level, either in arranging cards, e.g., as P44 (12P, 6R) said, "myself, I could make the organization make sense in my head, but [with HIRO] I wanted to make sure it was clear so they could detect the different clusters," or in sharing the workspace, e.g., as P23 (14P, 4R) said, "I had to make sure to keep my hands back because HIRO was trying to read." At a conceptual level, P37 (16P, 5R) worked to make sure that HIRO could interpret P37's clusters, reasoning that "I wanted the robot to implicitly recognize the themes … so this motivated me to have ones that are really similar together."

On the other hand, some participants felt that simply having HIRO place some cards reduced their workload. P36 (9P, 2R) reasoned that HIRO "basically [made] the same decisions I would have made. I could have had the same outcome without him but maybe it was easier on me and less decision fatigue." For some, working with HIRO reduced overthinking. "Alone … there are different aspects I would rethink over and over again," explained P20 (9P, 2R), "compared to if the robot would just take the card and place it in a particular stack [and] I would try to think of the why and not." For others, HIRO offered a form of automation, as per P1 (20P, 2R): "the robot just became more of a time-saving machine. It's, like, a way to double my bandwidth." In the end, P30 (13P, 5R) mused, "It comes down to, do you want to work on a group project by yourself? Or do you want to use it with somebody who is helpful to you and that you can try and figure it out together? I think just cognitively it helps you take a load off, right?"

6.4 How HIRO Affected the Pacing and Flow of the Task

In line with our quantitative findings, several participants noted that working with the robot slowed down the pace of the task. Some attributed this to the physical delay in waiting for the robot to find, pick, and place a card. As P11 (14P, 3R) described it, "You had to wait for the robot a lot. Sometimes, it would take a while to go and get the card and figure out where it wanted to put it." Alternatively, P2 (15P, 6R) explained that thinking things through with HIRO took more time: "If HIRO put a card down … that I was unsure of, I would think about it more deeply. Whereas when you're working alone, you're, like, this is my thought. So, that's why it's faster." In contrast, some participants actually found the robot to move too quickly, in a way that rushed them or disrupted their desired workflow. P15 (13P, 6R) told us, "When I work with the robot I feel a bit tense because I felt like I have to keep up with its pace."

Beyond pacing, some found that working with HIRO changed the flow of the task, forcing them to process the cards incrementally. "I think working alone was a bit easier because first I got to read all the cards before deciding what clusters to make and while working with the robot I had to decide as it happened, as we placed the cards, which clusters to create," explained P8 (13P, 3R).

While several participants preferred the faster pace when they worked alone, some participants found value working at a slower pace with the robot as it offered them more time to reflect on their decisions as well as to plan ahead as the robot was working. "Definitely, it takes more time to work with a robot, which is good because I think it allowed me to re-examine my decisions," said P16 (26P, 8R). "I had more time while the robot was working to conceptualize the different sections that were going on," said P40 (16P, 4R).

6.5 Enjoyment: The Experience of Working with HIRO

Participants found certain aspects of working with HIRO to be enjoyable in their own right. For some, working with a robot was simply more engaging. "I don't really know why that is, but it was fun to work with the robot," said P5 (15P, 4R). Others felt a sense of social connection to the robot. "It was friendlier," said P3 (14P, 1R). "I don't know how else to put it, but it was more enjoyable." "I feel like we connected. I feel like we're friends. I would be very sad if HIRO doesn't like me," said P52 (16P, 8R). Finally, for some, working with HIRO took on an element of mystery, and trying to understand HIRO added an interesting dimension to the task. "I would use it sort of like a game. I would see how it interacted. It would be more fun for me to interact in that way," said P53 (18P, 2R).

6.6 Nonverbal Communication: Limits in Creative Collaboration

HIRO was designed to communicate with participants through placements in the shared affinity diagram and had no ability to communicate verbally. To some, the lack of verbal communication made the mechanics of collaboration more challenging, specifically wanting support for explanations and debate.

Participants desired to both explain their reasoning to HIRO and hear HIRO explain its reasoning to them. "I can't talk to the robot and explain my thinking," said P8 (13P, 3R). "And if HIRO was a human, I would ask them what they were thinking." P7 (14P, 7R) put it a bit more bluntly, saying, "I don't know its thoughts so I cannot brainstorm with it." HIRO's inability to communicate the rationale behind its choices led some to reject it as a collaborator. "I feel like because HIRO can't express their opinion," P21 (14P, 2R) told us, "I can't really come to a compromise, so I decided to just go with what I thought. So it's not very much a collaboration."

Participants felt that verbal communication would allow for more argumentation between human and robot. "With a person," reasoned P4 (12P, 1R), "you'd be constantly bouncing ideas back and forth and trying to come to some sort of middle ground that you both agree on," whereas, as P37 (16P, 5R) put it, "The robot is just sort of saying, 'Here,' and you can take it or leave it." This placed the burden of handling disagreements on the human. P17 (11P, 5R) explained, "[humans] can actually exchange information if there's a disagreement or if any modifications need to be made . . . but with HIRO, it's me handling the situation." Overall, P18 (21P, 12R) argued, "a collaborator should demonstrate its argument."

Participants, however, were split on whether they would rather work with an agent that could argue with them or not. Some participants appreciated the ability to overrule HIRO without having to argue with it or account for its feelings. As P9 (14P, 3R) told us, "From my experience working with other groups, there must be a compromise when there's idea conflicts. And, for me, it's very inefficient during the working process. But for the robot . . . I noticed that we do not have conflicts." Some described HIRO's muteness as a good fit for their personality and collaboration preferences. As P36 (9P, 2R) put it, "I'm a little introverted and he doesn't talk back but he gets the job done." That said, silence could be socially jarring, e.g., as P12 (15P, 6R) put it, "It feels like my thoughts, my worries, my insecurities are just put out there and I just don't have anyone to reassure me or guide me in the right direction."

6.7 Perceived Intelligence: Can a Robot Understand Human Issues?

Participants' mental models of HIRO could color their interactions with it. Participants rated HIRO as relatively intelligent (see Figure 13). However, assumptions about how HIRO was processing information shaped expectations about how it could assist on this task. For example, some differentiated the levels of insight at which an algorithm and a human could characterize human needs. "I think it went about as I expected it to go, a robot sorting human concerns," said P5 (15P, 4R), "people obviously innately understand human struggles more than many robots are capable of." Occasionally, HIRO's decisions challenged such assumptions. P12 (15P, 6R), for example, believed that, "this context as a human that you have surrounding each of these comments . . . HIRO might not have that sort of context." However, they told us, as the study went on, "a lot of the choices [HIRO] made, I could reason myself and get to the same place he did. So, there was more thought I guess, and less algorithm."

HIRO's movements also influenced participants' perception of its intelligence and agency, for example, the way that HIRO scanned the diagram before placing a new card. "It was almost like watching someone and sort of following their thought process," said P22 (7P, 2R). P12 (15P, 6R) suggested, "It's very mechanical, but also I could see a human quality kind of like pondering as he went." "Sometimes it feels as though it's fast. Other times it feels like it's slow," said P21 (14P, 2R), "sometimes it feels enthusiastic. Other times, it's like taking time to process and be careful." P36 (9P, 2R) interpreted HIRO's movements in a collaborative sense, "like he was taking a second to think about how I was thinking before placing his card." Not everybody felt the same way. P1 (20P, 2R) and P31 (7P, 2R) expressed confusion as to why HIRO was moving over the diagram or moving in between turns, respectively, whereas P18 (21P, 12R) found it distracting if HIRO moved when P18 was trying to think.

6.8 Roles and Power Dynamics: Making Creative Decisions with HIRO

Participants attributed different collaborative roles to HIRO. Some viewed HIRO as a reference tool to check their thinking or to offer ideas when needed. For example, P1 (20P, 2R) said that HIRO, "gave me more possible solutions or options around which way I should go with organizing." Others allowed HIRO to assume more control, including delegating clustering decisions. P9 (14P, 3R) explained that, "when I make sure that it's in the same logic with me, there's no need for me to carefully read every card. And it's automatic." At an extreme, P43 (13P, 0R) told us that if given the chance to work with HIRO again, P43 would let HIRO "place everything and just play on my phone, and then after HIRO does it, I can just check its work."

Oftentimes, roles split across defining categories and sorting cards. P23 (14P, 4R) described HIRO as being "constrained to what I had set forth," and "I was the sole person making categories, then [HIRO is] somebody who's just helping me sort things." Others simply preferred the role of creating categories. As P27 (12P, 3R) said, "I don't want the robot to start making new categories for me, because I kind of wanted to make the categories." Some participants, however, used HIRO's input to define categories, e.g., P5 (15P, 4R): "[HIRO] would put things down and instead of having to generate categories myself I could either agree or disagree with the robot."

Sometimes, roles reflected power dynamics with HIRO. P16 (26P, 8R) described deferring to HIRO: "when I first started doing the card sorting, I definitely saw our power dynamic; the robot had more power and knowledge in this area." For P15 (13P, 6R), power imbalance inhibited collaboration: "I think it's smart. So I don't want to challenge it. Sometimes I think it places the card in the wrong place, but I'm not sure if I should move it. So I prefer working alone." Others felt more comfortable overriding HIRO as they saw fit. "I think my reasoning trumps his," P12 (15P, 6R) explained, "but not in a disrespectful way." In the end, some participants observed that HIRO's limitations, physical or otherwise, gave more power to the human. "I think ultimately the human will always have at least just a little bit more control," P37 (16P, 5R) told us, "because the robot can move it so many times but I can also just move it . . . a lot quicker than the robot." Sometimes, this came with emotional consequences. P52 (16P, 8R) worried about hurting HIRO by rejecting its suggestions: "It felt like I was offending him, because it can't defend itself, right? It can't be like, 'No, actually, I'm going to put it back. So I'm kind of the decision power here."

7 DISCUSSION

We have described a system and study that we developed to research human–robot collaboration on a sensemaking design task, specifically, need-finding through an affinity diagram. We were interested in three questions: how working with HIRO might influence how humans construct affinity diagrams, how affinity diagrams perform as a human–robot collaborative medium, and what to consider when designing robots to support activities such as affinity diagramming. We start by discussing the first two of these questions, followed by a consideration of the unique context of our participants' limited experience with affinity diagramming. We then present a set of implications for the future design of human–robot collaborative design systems.

7.1 How Did Working With a Robot Influence How Humans Constructed Affinity Diagrams?

Our findings suggest several ways that HIRO's presence affected participants' affinity diagramming experiences. Participants tended to work more slowly with HIRO. While this could be attributed to the speed of the system or the overhead of collaboration, it is worth noting that some participants spent *less* time on the task when working with HIRO (see Figure 8). Further, despite the turn-taking dynamic, the division of time between human and robot effort was not a zero-sum game: several participants described thinking about or working on the task while HIRO was moving. While they did not report overall differences in cognitive load, several participants described ways that HIRO affected their cognitive load, stress, enjoyment, or social engagement. Working with HIRO could inspire participants to consider alternative interpretations of the data. Sometimes, this increased perceived effort and the difficulty interpreting choices could be frustrating. HIRO could also mitigate uncertainty about individual cards or the overall direction of the diagram.

These findings reflect prior studies on how interaction with other humans affects cognition and creativity. Tversky and Hand [67], for example, found that the mere presence of a human actor in a photograph encouraged people to adopt the actor's perspective when describing objects in the photo. Paulus cites findings that suggest that exposure to others' perspectives can increase individual creativity [57], affording a wider base to explore ideas, commingling cognitive styles, and applying heterogeneous knowledge sets to a problem. At the same time, group social dynamics such as productivity blocking, evaluation apprehension, and free-riding can inhibit creativity [16].

That said, while participants saw HIRO as intelligent and likeable, they tended not to perceive it as anthropomorphic. Some participants believed that a robot could never understand human concerns, and participants were frustrated that HIRO was unable to explain and defend its ideas like a human. This mirrors the argument by Guckelsberger et al. that creative agency in machines requires not just creative acts but also explanations that reflect and maintain a creative identity [26]. This remains a significant challenge for human–machine collaborative design.

7.2 How Might Affinity Diagrams Support Creative Collaboration Between a Human and a Robot?

The context we investigated here, affinity diagramming for need-finding, is unexplored in the human–robot creative collaboration realm. How did the medium of shared note sorting perform in the design process?

As in human collaboration in affinity diagramming, participants usually used spatial position to communicate with the robot about the relationships between notes. Even when they disagreed with HIRO, participants attributed conceptual meaning to where the robot placed the cards. In many cases, they constructed mutual understanding with HIRO about how to interpret the data over the course of several card placements.

That said, we observed clear communication limitations to collaborating through the diagram alone. In a few cases, participants were confused by ambiguous card placements. Participants also disliked the unequivocal nature of HIRO wordlessly placing cards, wanting to hear its reasoning or to negotiate with it. This, combined with uncertainty around how HIRO was thinking, could restrict the degree to which participants felt like they were collaborating well with the robot. In short, the affinity diagram was sometimes effective at communicating *opinions*, but not *reasoning*, and HIRO's movements over the diagram expressed *that* it was thinking but not *what* it was thinking.

Ultimately, while the shared affinity diagram did not consistently fulfill participants' collaboration needs with HIRO, we also saw glimpses of how such a diagram could serve as the engine of mutual understanding in a highly unstructured task, such as need-finding.

7.3 Interpreting Participant Experiences Through a Lens of Task Expertise

On reflection, several elements of the themes we observed reflect novice design behaviors that may have emanated from our participants' relatively low self-reported experience with affinity diagramming.

Novices and experts are known to exhibit different behaviors, and much work has gone into characterizing these differences in the context of unstructured problems. Cross's comprehensive survey of expertise in design [12] finds several important tendencies that characterize experts in early-stage design work: experts are solution focused and use conjectures to scope problems early on; experts tend to adopt to and stick to early design concepts rather than exploring many alternatives; and, finally, experts are opportunistic in their methods and frequently switch between parallel cognitive activities.

The divide between the tendencies that Cross attributes to expert designers versus novice designers surfaced in our participants' stories and the themes we extracted from our interviews. For instance, several participants remarked that HIRO gave them a direction to explore or injected objectivity into the task, suggesting a lack of willingness or ability to conjecture from experience. Participants also noted that working with HIRO required a collaborative overhead compared with working alone. This could be distracting, preventing participants from following "the first thing that popped into [their] head" or applying a more top-down structured process. In contrast, others noticed themselves exploring different options or getting un-stuck by watching HIRO make a decision, evoking the kind of cognitive switching that an expert might already use effectively in unstructured problem-solving.

Experts who were more familiar with this particular design activity may have had very different experiences working with HIRO than our participants. For example, switching between different ideas or ways of thinking could be more distracting for an expert who is already thinking along parallel tracks. Likewise, using HIRO as a proxy for conjecture and problem-scoping would likely not have appealed to an expert with the accumulated experience to make those decisions confidently and intuitively. While HIRO making choices could offer a safe sense of direction to someone who feels uncertain, the same behaviors could be burdensome for an expert user who would prefer control and flexibility to opportunistically define their own path. On the other hand, an expert's solution-oriented focus might provide more useful context to the robot compared with a novice more fixated on the present.

In sum, we might expect experts at tasks such as affinity diagramming to look for different kinds of help than novices would in ways that are more tailored to their own processes rather than broadly suited to the general design activity at hand.

7.4 How Should We Design Robots to Support Conceptual Aspects of Designing?

Based on what we learned about affinity diagramming user needs with a robot, we propose the following guidelines for designing robots to support similar conceptual design activities, particularly with novice users.

(1) **Account for the robot's speed.** Consider how fast a robot moves when determining the roles it plays to support design activity. For various reasons, participants in our study tended to spend more time on their diagrams when working with HIRO than when they were working alone. One factor that contributed to this was the perceived need to wait for HIRO while it was making a move. For at least one participant, the need to take turns

stifled the individual's preferred flow for running through all the cards before choosing placements. Others, in contrast, suggested it moved too quickly, disrupting them or making them feel tense. Our findings suggest that the perceived effects of a collaborative robot's speed may influence the flow of a creative activity in nuanced ways. By conventional human–robot fluency metrics such as idle time [33], a slow robot should be judicious about undertaking tasks with long-running motions. That said, idle time can also offer a human partner time to stop and think while sharing the floor. For some of our participants, the difference in pacing from working alone provided opportunities to re-examine their choices or conceptualize the meaning of the current clusters. Overall, for different participants and purposes, the speed at which a robot moves could have both benefits and challenges in terms of cognitive collaboration.

This suggests that the speed at which a robot is designed to move should align with the needs of a specific human collaborator and task. Of course, technical or safety limits may constrain the degree to which a robot's speed can be adjusted. With this in mind, an alternative approach is to consider the kinds of tasks the speed at which a robot can move are suited for.

This guideline connects to the different roles participants gave the robot. For instance, a slower robot might be better suited to define clusters or handle difficult cases, whereas a faster robot might be more appropriate for automation or tasks that can be completed asynchronously without the human's attention.

(2) **Pursue mutual understanding in creative collaborations.** Second, consider the system design goal of mutual understanding rather than more straightforward goals such as agreeing with the user or exposing the user to alternatives. Getting on the same page with HIRO was a process that informed many participants' experiences. One participant described the experience as building up to speaking a common language with HIRO, something only possible once some critical mass of shared experience had been achieved. Participants' willingness to engage in this process suggests several opportunities, most notably for novices at an unstructured task.

For example, establishing mutual understanding with a robotic partner could provide a gentle entry point for exploration. Our interviews revealed that some participants credited HIRO with providing early direction or a sense of objectivity. That said, simply providing initial directions could inhibit a novice's willingness to explore more deeply [9]. The process of developing mutual understanding with HIRO suggests a better alternative. Many participants described testing, rationalizing, and rebutting HIRO's choices, with a desire to engage them more discoursively. If the system prioritizes this process of developing mutual understanding with a tool such as HIRO, it could offer low-stakes incentives to nontrivially engage in speculative directions for those without the confidence that comes with expert intuition.

What might this look like? Bratman describes shared cooperative activity as rooted in mutual responsiveness, demonstrated commitment to the joint activity, and commitment to mutual support [10]. Through utterances or actions, a robot might might explicitly signal its commitment to developing mutual understanding with a user beyond simply finishing the task. For a robot such as HIRO, actions such as frequently inspecting the human's clusters before making a move or identifying and pointing out clusters that it finds less cohesive might encourage the human to reciprocate and, ultimately, support both more collaboration and exploration.

(3) **Identify opportunities for constructive disagreements.** Beyond the dynamics of developing mutual understanding between a human and a robot, our findings suggest that

there may be opportunities for a robot to challenge humans in a collaborative sensemaking task. Several participants described needing to think more about their choices or changing their mind when working with HIRO. One of our participants described needing more *surprise* from HIRO to see it as a compelling creative partner. This resonates with prior work in computational creativity support, including several projects that suggest ways to encourage creative shifts (e.g., [42, 54]) or explore formalizing surprising ideas [24], pointing to the usefulness of constructive disagreement.

That said, not all disagreements are equal: our participants distinguished between disagreements they could rationalize and ones that didn't make sense to them. In some cases, the content of the disagreement could be immaterial: one participant told us that simply seeing a decision from HIRO helped the individual to reach one's own, regardless of agreeing or not. In short, characterizing the role of disagreement in a creative collaboration with a robot is multifaceted.

Unsurprisingly, this mirrors the complexity of conflict in human creative collaborations. Badke-Shaub et al. found that design teams that were relatively more confrontational and less collaborative tended to generate more functional and innovative ideas, although they still exhibited mostly collaborative behavior [5]. However, conceptual conflict can escalate to damaging affective conflict in teams [3, 64], a relevant design consideration for any collaborative robot with social behaviors. We saw elements of this in participants who perceived a social connection or power dynamics with HIRO that affected their willingness to override it when they disagreed with HIRO. This kind of dynamic could present both a constraint and a degree of opportunity to push creative boundaries within reason, using the social connection as a kind of buffer to affective conflict. Others appreciated that HIRO was non-confrontational, offering a second perspective without apparent social consequences. Finding the right balance of creative agreement and confrontation between a human and robot in an unstructured task demands attention to what constitutes constructive or destructive disagreements in each context.

The context of physical collaboration with a robot adds a particularly salient dimension to this discussion. Klemmer argues that embodiment carries risk, because choices are more visible and physical actions express commitment to those choices [45]. This sense of risk is an interesting lens through which to view our findings around second perspectives and power dynamics, whether participants felt hesitant to override HIRO's choices or simply anxious feeling their own actions were being scrutinized. Klemmer et al. point out that this riskiness can work both ways: increasing focus and attention on the task but also constraining willingness to think divergently. Beyond how participants perceived risk in their own actions, the visibility of HIRO's actions also carried risk in terms of how participants perceived its intentions, intelligence, or helpfulness. In literally changing a participant's arrangement of cards, HIRO expressed a commitment to a particular interpretation of the data that demanded a response. In a sense, the purely physical and uncompromising nature of HIRO's behaviors amplified this risk—it had no way of, for example, lowering the level of commitment by verbally expressing uncertainty about a placement. This leads into our final design guideline, which addresses the nature and limits of communicating solely through shared physical materials.

(4) **Use other modalities of communication in conjunction with physical materials.** Participants' desire to converse with HIRO reflected limitations on what HIRO was able to communicate nonverbally. While the diagram communicated relationships between notes, participants wanted to discuss the motivations behind choices that they or the robot made.

To support this, we suggest using physical representations such as affinity diagrams in conjunction with other modalities that can add depth when needed.

While verbal dialog that includes explanations and debate is a straightforward way to add nuance, this is not the only solution. For example, one participant suggested highlighting key text on cards. HIRO could also use gestures to broaden communication; Heiser et al. describe teammates gesturing over maps to support collaboration and cognition [32]. Some of our participants gestured with or over note cards, and many interpreted meaning in HIRO's movements over the diagram. Beyond pointing gestures, the functions of metaphoric and iconic gestures in collaboration and creativity have been studied in human teams [53, 71]. While less explored in HRI (e.g., [36]), metaphoric and iconic gestures may be particularly useful to ground creative exploration in human–robot collaborative design.

A physically shared diagram may afford forms of communication beyond the intended conceptual organization of the diagram itself. For example, while coding placements, we observed participants covering a cluster to hide it from HIRO or placing cards between clusters to indicate indecision. Overall, there is a rich interaction space to explore at the boundaries of what material representations explictly afford, a space that might yield more intricate human–robot dialogues of the sort that arise in unstructured problem-solving.

8 LIMITATIONS AND FUTURE WORK

The study and findings that we have described have several important limitations with respect to their scope and generalizability.

8.1 Study Limitations

In terms of scope, our hypotheses and study measures were designed to provide a starting point to characterize what working with a robot on a sensemaking task such as affinity diagramming might look like. In combination with our qualitative analysis, we hope this provides a foundation for future exploration. However, given the rich nuances that we observed, we caution against interpreting broad measures such as completion time in this context normatively. It is also important to consider the relative inexperience of our participant pool with the design activity and how this reflected in our findings. As we've discussed, novices and experts are thought to behave quite differently in unstructured activities, and our observations of our participants often aligned with tendencies of novice designers. We thus caution against applying our findings and takeaways to expert designers. Nonetheless, we believe that this work provides useful insights for the design of robotic systems to specifically support novices at unstructured design activities such as affinity diagramming.

Another limitation of this study concerns the limited data that participants diagrammed. For the sake of time and, more importantly, the limitations of HIRO's workspace, participants only organized 28 user notes. In reality, most affinity diagrams are constructed to parse many more notes than this. At least one participant pointed out that scaling the robot would be an issue, and it is not clear how scaling up the data and workspace might influence how people collaborate with the robot.

Finally, despite our efforts to standardize the treatments, some participants noticed a learning effect between treatments or felt that one set of user data was easier to work with than the other. While we randomly counterbalanced our treatments to offset this, the individual experiences we recounted could still have been influenced by this effect.

8.2 System Improvements and Future Directions

While it was not the main focus of this study, HIRO's semantic encoding could likely be improved with several simple changes, including evaluating different distance metrics and performing dimensionality reduction before clustering.

Additionally, to simplify its behavior for this study, HIRO only added cards to the diagram. It is fairly straightforward to imagine strategies of, for example, splitting large clusters by subclustering the spatially informed embeddings obtained for the notes in that cluster. While increasing the variety of actions that HIRO takes should expand the ways that it can support a human user, it also introduces a number of design challenges in determining when it should take what behavior, and how the behaviors should be implemented, considering the guidelines we posed above.

9 CONCLUSION

HIRO is a tabletop robotic arm designed to affinity diagram textual data with a human partner, developed to study what it might look like for a robot to engage with a human in a sensemaking design activity. In a within-user study of n = 56 novice participants, we found evidence that working with HIRO increased the time that participants spent on the task without compelling evidence for any corresponding effect on cognitive load. Post-study interviews attributed the temporal increase to time spent accounting for both the speed of the robot and its point of view. Individual participants suggested aspects of working with the robot that were more or less cognitively demanding, depending on how they chose to collaborate with it. This, in turn, varied according to a combination of how participants perceived HIRO's intelligence and how they interpreted power dynamics between themselves and the robot. Participants sought mutual understanding with HIRO about the task and each other. This developed over time alongside their own understanding of the shared data. These efforts also revealed limitations to nonverbal communication in supporting human–robot collaborative sensemaking.

Based on these findings, we suggest that those who develop robots to collaborate with human designers on sensemaking tasks design around the robot's speed, frame the collaboration to pursue mutual understanding rather than only design outcomes, identify opportunities for constructive disagreements, and use verbal and nonverbal modalities of communication in conjunction with physical materials.

APPENDICES

A TASK COMPLETION TIME ANALYSIS DETAILS

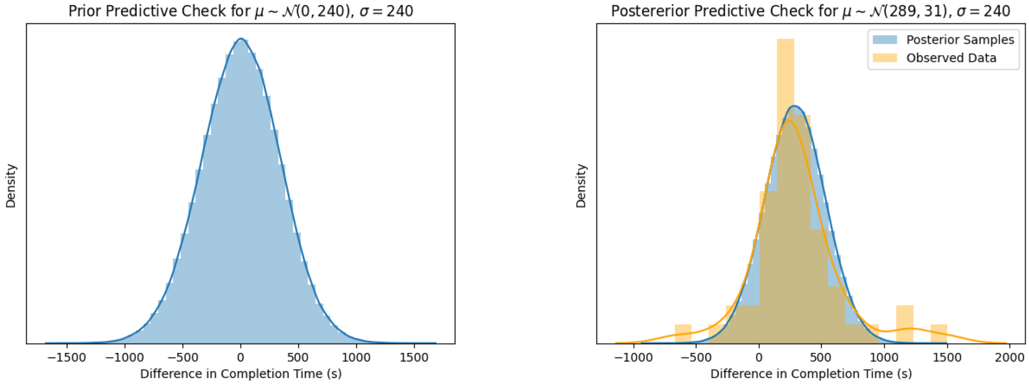
The findings on **H1** were based on a Bayesian analysis of the recorded differences in completion time between treatments. This difference was modeled as a normal distribution with an unknown mean and fixed standard deviation of 240 seconds.

A.1 Prior Predictive Checks

For our analysis, we adopted a weakly informative prior over the mean as a normal distribution centered around zero with a standard deviation of 240 seconds. The prior predictive check in Figure 14(a) shows 1,000 samples generated from a normal distribution with a fixed standard deviation of 240 for each of 1,000 draws from this prior over μ. The outputs are reasonable, with most of the density between −1,000 and 1,000 seconds, or roughly 17 minutes, which would be a very large but not unreasonable time difference.

A.2 Posterior Predictive Checks

To run a posterior predictive check, we sampled 1,000 values for each of 1,000 draws of μ from the posterior distribution. The resulting samples are plotted in Figure 14(b) and resemble the overlaid observed data.



(a) Prior predictive check for H1.

(b) Posterior predictive check for H1.

Fig. 14. Prior and posterior predictive checks for H1.

A.3 Sensitivity Analysis

To test the sensitivity of this result to our choice of prior, we calculated the posterior using different parameterizations of the mean prior ($M_\mu \in \{-120, -60, 0, 60, 120\}$ and $S_\mu \in \{180, 240, 300, 360, 420, 480, 540\}$) and different values for the fixed standard deviation ($S_y \in \{180, 240, 300, 360, 420, 480, 540\}$). We also tested a Half-Cauchy prior over S_y instead of fixed values, with varying scale parameters ($\beta \in \{120, 240, 360\}$), using MCMC via PYMC.[1] The HDI and mean of each posterior for μ with fixed S_y is plotted in Figure 15 and with the Half-Cauchy priors over S_y in Figure 16. Across all tested priors, the HDI is above the ROPE, without overlap, suggesting that the observed effect is robust to our choice of priors.

[1] Thomas Wiecki, John Salvatier, Ricardo Vieira, Maxim Kochurov, Anand Patil, Michael Osthege, Brandon T. Willard, Bill Engels, Colin Carroll, Osvaldo A. Martin, Adrian Seyboldt, Austin Rochford, Luciano Paz, rpgoldman, Kyle Meyer, Peadar Coyle, Oriol Abril-Pla, Marco Edward Gorelli, Ravin Kumar, Junpeng Lao, Virgile Andreani, Taku Yoshioka, George Ho, Thomas Kluyver, Kyle Beauchamp, Alexandre Andorra, Demetri Pananos, Eelke Spaak, and Benjamin Edwards. 2024. pymc-devs/pymc: v5.10.4. Zenodo. DOI: https://doi.org/10.5281/zenodo.10656993

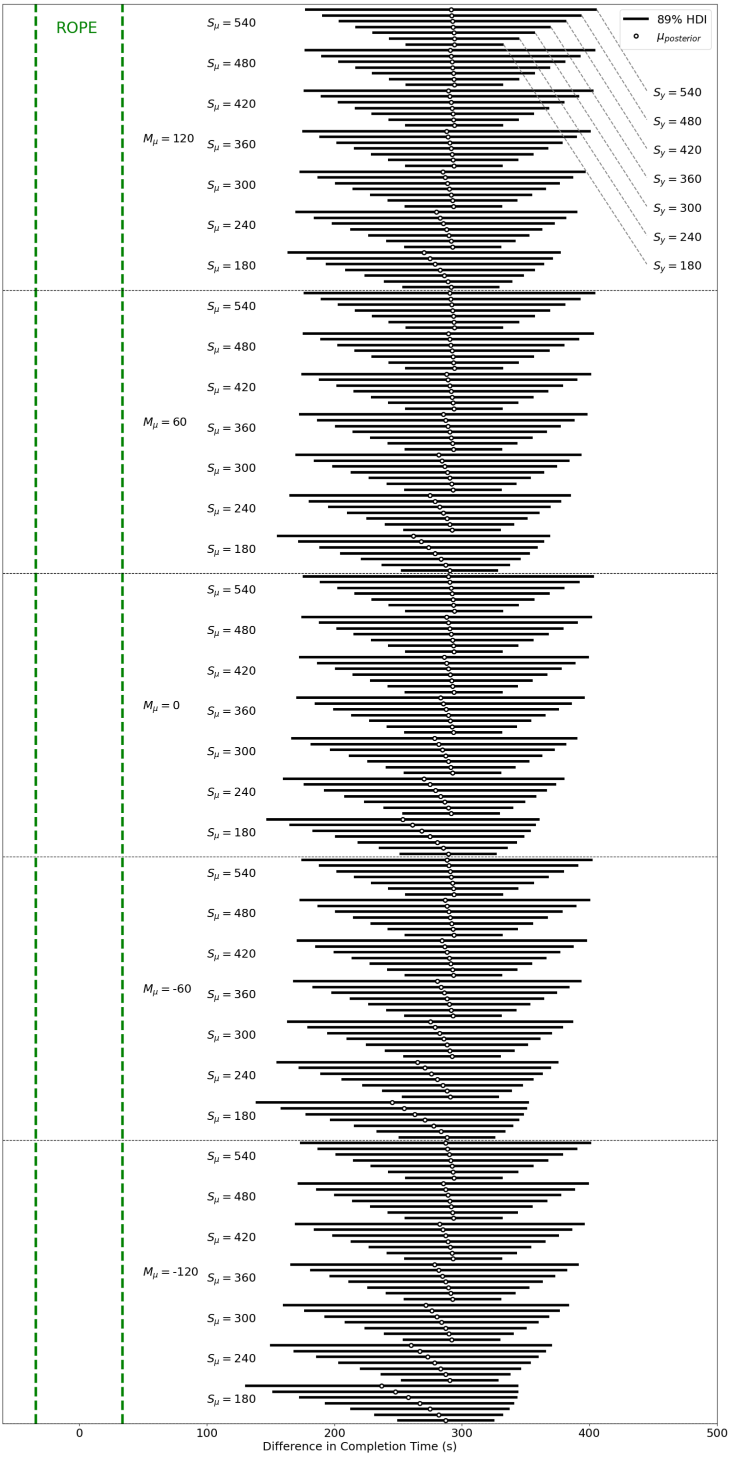


Fig. 15. Sensitivity analysis for H1 with fixed priors over S_y.

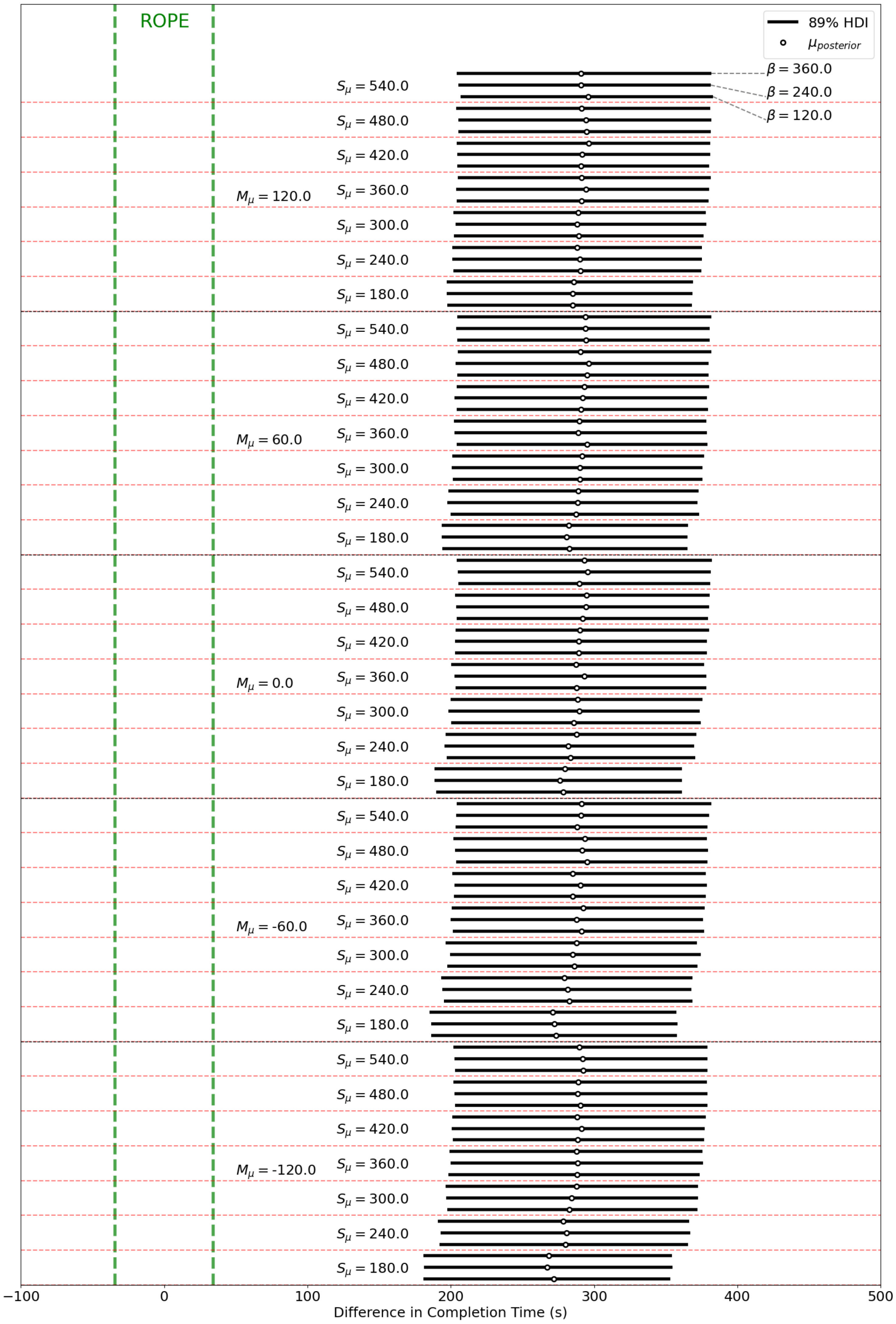


Fig. 16. Sensitivity analysis for H1 with Half-Cauchy priors over S_y.

B COGNITIVE LOAD ANALYSIS DETAILS

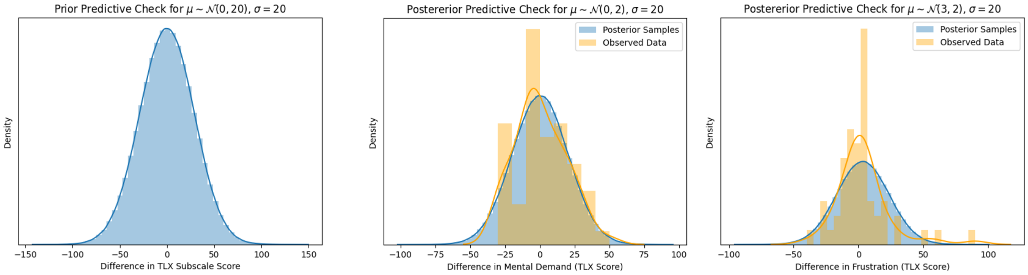
For the two findings regarding **H2** and **H3**, we performed a Bayesian analysis of the reported differences in mental demand and frustration, respectively, via the corresponding subscales of NASA TLX. We modeled both of these differences as normal distributions with an unknown mean and fixed the standard deviation at 20 points based on the underlying 100-point scale.

B.1 Prior Predictive Checks

For both subscales, we adopted a weakly informative prior over the mean of a normal distribution centered around zero with a standard deviation of 20 points' difference. The prior predictive check in Figure 17(a) shows 1,000 samples generated from a normal distribution with a fixed standard deviation of 20 for each of 1,000 draws from this prior over μ. The density of the outputs falls almost entirely between -100 and 100 points, the maximum possible differences for each TLX subscale, with most of the density between -50 and 50.

B.2 Posterior Predictive Checks

For each of the posterior distributions, we sampled 1,000 values for each of 1,000 draws of μ from the posterior distribution. The resulting samples, plotted in Figures 17(b) and 17(c) mostly resemble the distribution of the overlaid observed data, although the kernel density estimate of the frustration data has a somewhat higher peak and thicker right tail than the distribution of samples generated from draws from the posterior.



(a) Prior predictive check H2, H3. (b) Posterior predictive check H2. (c) Posterior predictive check H3.

Fig. 17. Prior and posterior predictive checks for H2 and H3.

B.3 Sensitivity Analyses

To gauge the sensitivity of our results to our choice of prior, we calculated the posterior for each subscale with their respective observations using different parameterizations of the mean prior ($M_\mu \in \{-20, -10, 0, 10, 20\}$ and $S_\mu \in \{5, 10, 20, 30, 40, 50\}$) and different values for the fixed standard deviation. We also tested a Half-Cauchy prior over S_y instead of fixed values, with varying scale parameters ($\beta \in \{5, 20, 50\}$) using MCMC via PYMC.[1] The HDI and mean of each posterior for μ are plotted in Figures 18 and 19 and Figures 20 and 21 for H2 and H3, respectively. For mental demand, the HDI overlaps the ROPE in all trials except some with a fixed standard deviation, where $S_\mu = 5$ and $M_\mu = 20$ or -20, where it is above and below, respectively. For frustration, again, in some trials with a fixed standard deviation, where $S_\mu = 5$ and $M_\mu = 20$ or -20, the HDI falls outside the ROPE, above or below, respectively. Additionally, the HDI is consistently slightly above and outside the ROPE in cases in which the fixed standard deviation $S_y = 5$, which may not be a realistic prior insofar as the observed differences in frustration scores had a standard deviation of 20.4. In all other trials, including all of the trials with a Half-Cauchy prior over the standard

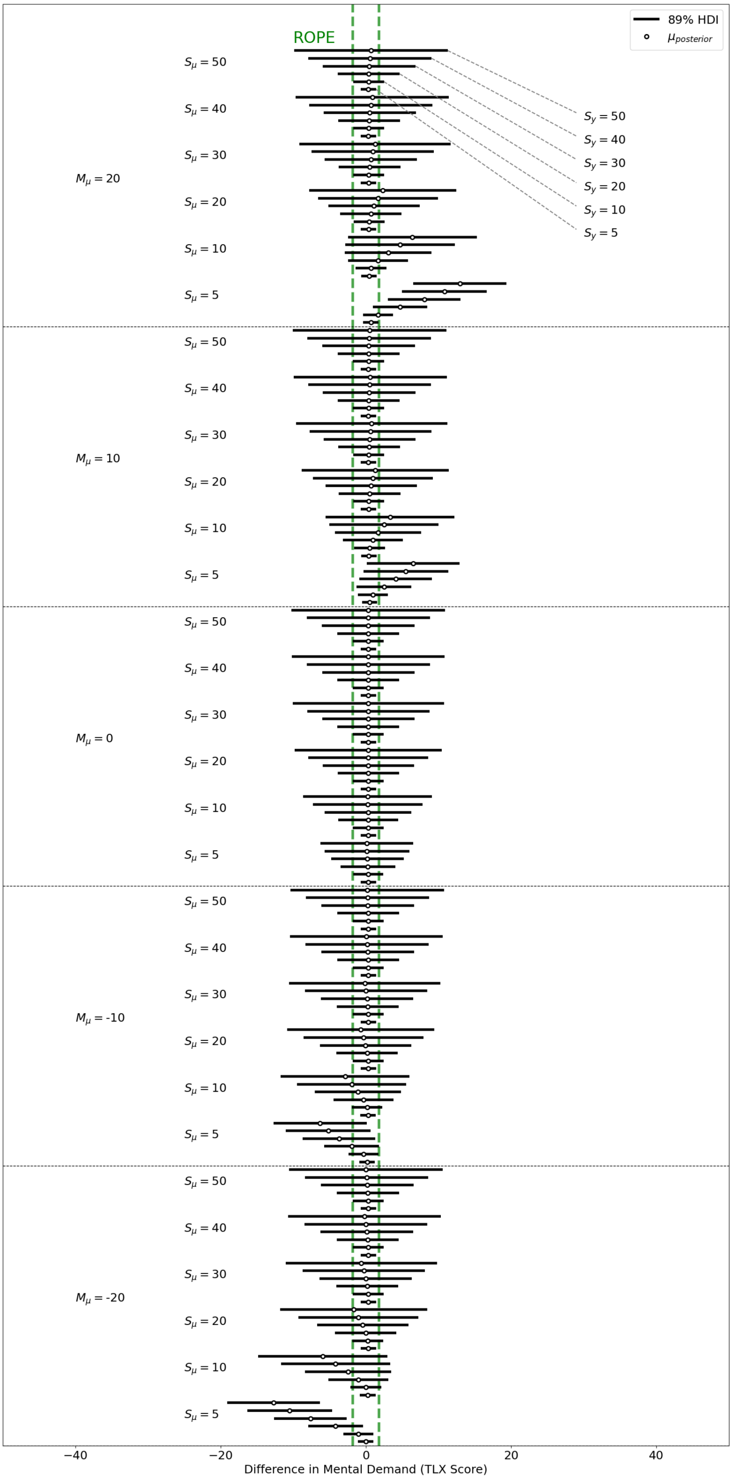


Fig. 18. Sensitivity analysis for H2 with fixed priors over S_y.

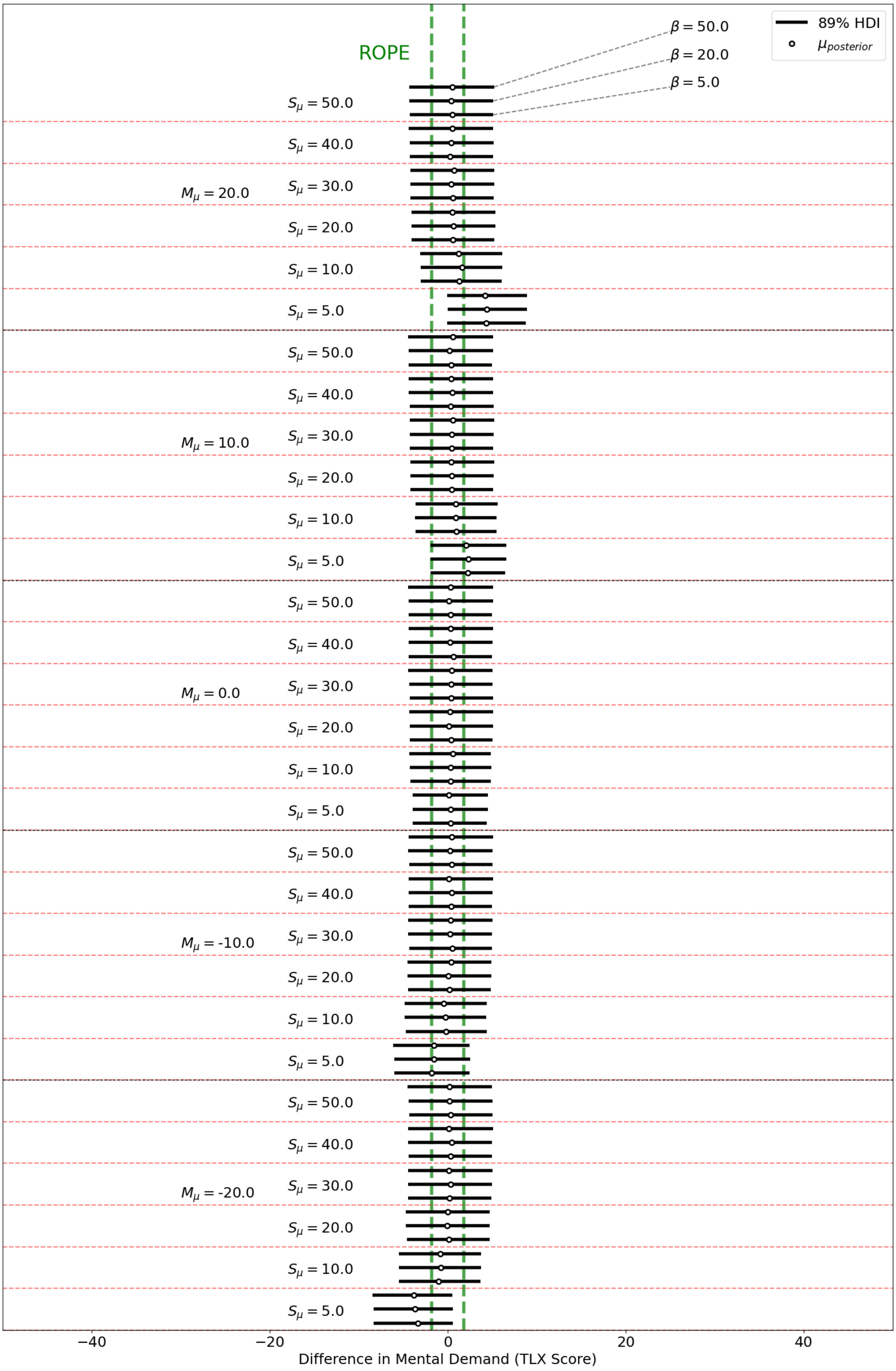


Fig. 19. Sensitivity analysis for H2 with Half-Cauchy priors over S_y.

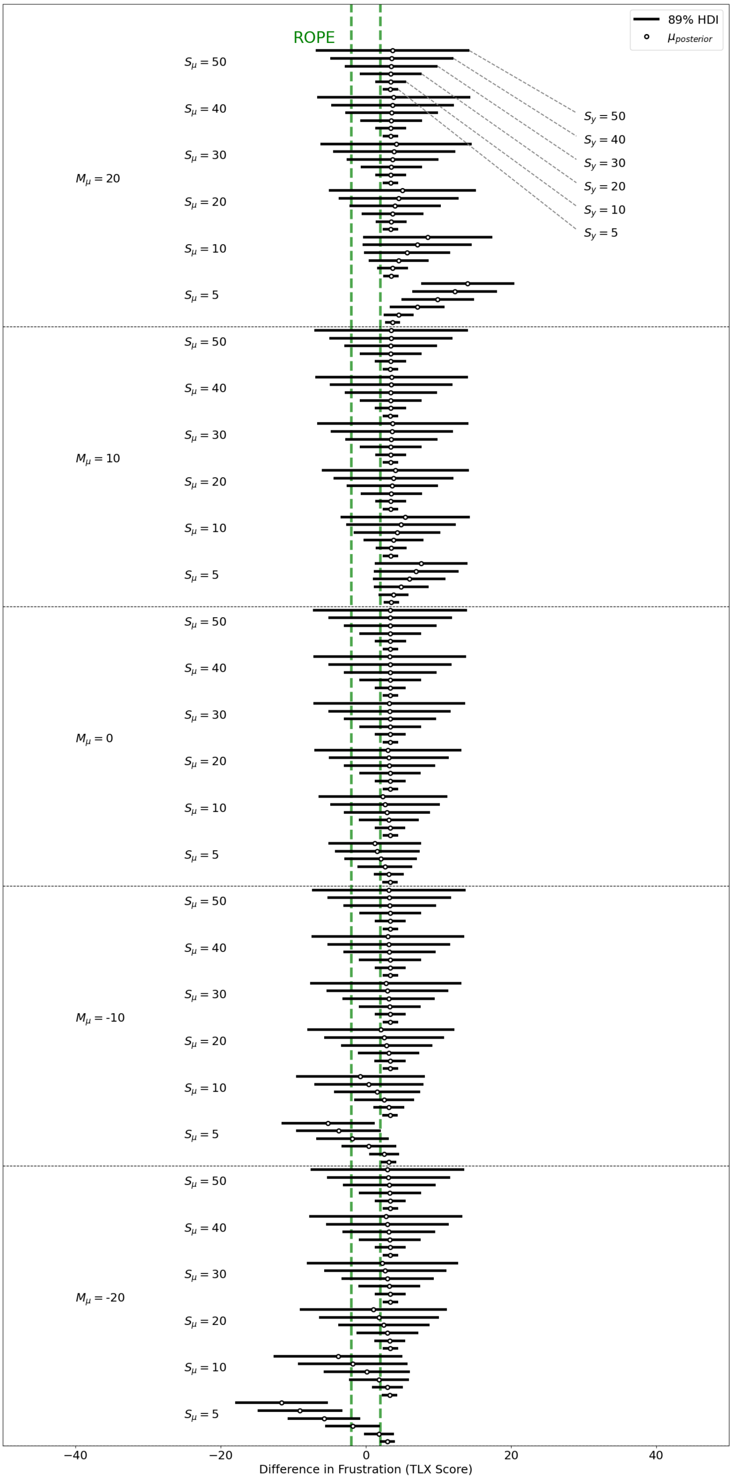


Fig. 20. Sensitivity analysis for H3 with fixed priors over S_y.

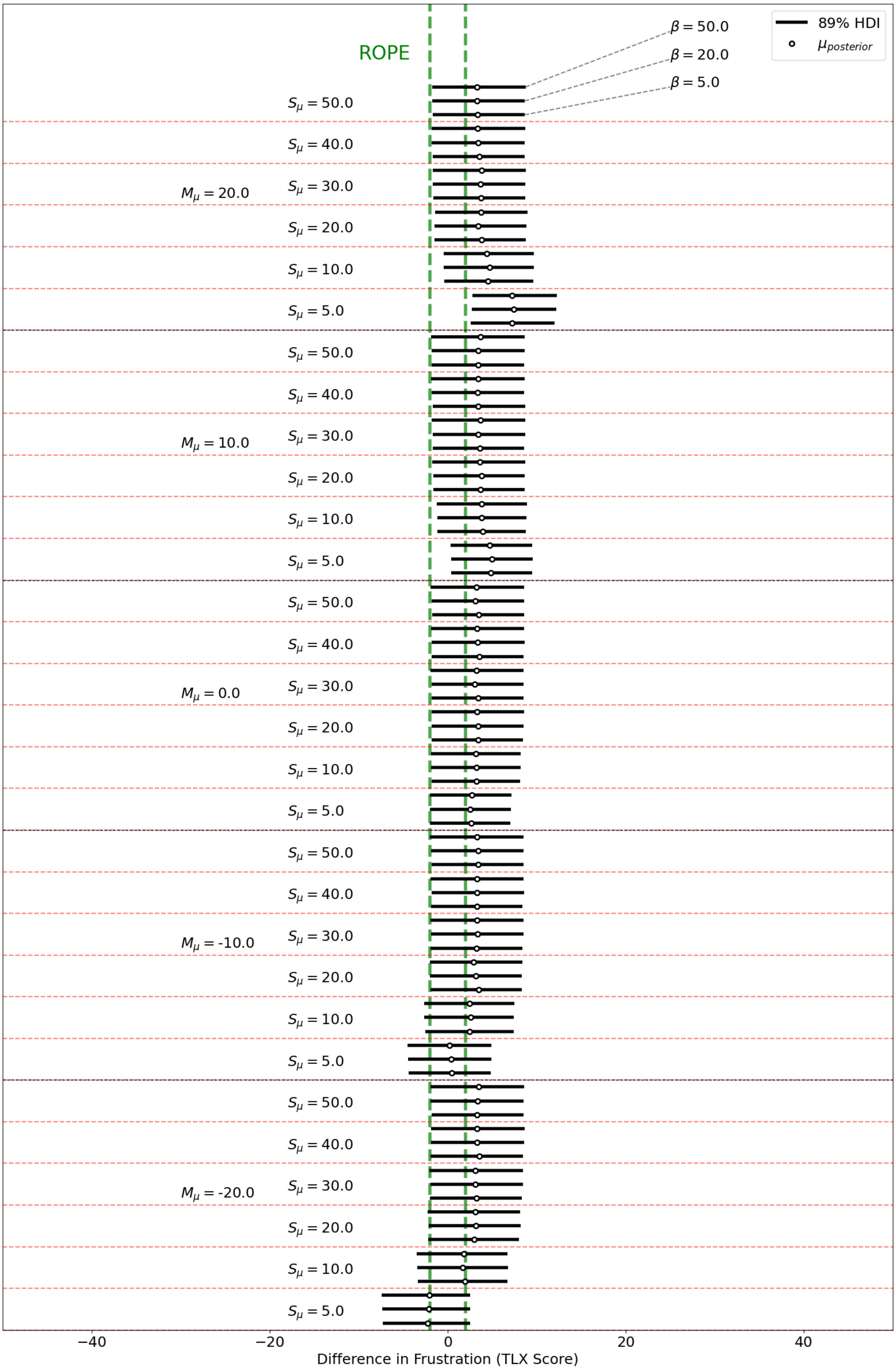


Fig. 21. Sensitivity analysis for H3 with Half-Cauchy priors over S_y.

deviation, the HDI overlaps the ROPE. Overall, although there are some cases in which the HDI is outside the ROPE, it would be difficult to conclude that there is a clear effect across choice of priors in either case, and certainly not the hypothesized effects of increased mental demand and reduced frustration.

ACKNOWLEDGMENTS

The authors would like to thank Adele Kong and Hamza Akik for their contributions to this project, as well as Matt Thomas and the Cornell Statistical Consulting Unit for their advice on experimental design, analysis, and reporting.

REFERENCES

[1] William C. Adams. 2015. Conducting semi-structured interviews. In *Handbook of Practical Program Evaluation* (2015), 492–505.

[2] Eman Abdullah AlOmar, Wajdi Aljedaani, Murtaza Tamjeed, Mohamed Wiem Mkaouer, and Yasmine N. El-Glaly. 2021. Finding the needle in a haystack: On the automatic identification of accessibility user reviews. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[3] Ofer Arazy, Lisa Yeo, and Oded Nov. 2013. Stay on the Wikipedia task: When task-related disagreements slip into personal and procedural conflicts. *Journal of the American Society for Information Science and Technology* 64, 8 (2013), 1634–1648. https://doi.org/10.1002/asi.22869 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22869

[4] Pranjal Awasthi, Maria Balcan, and Konstantin Voevodski. 2014. Local algorithms for interactive clustering. In *International Conference on Machine Learning*. PMLR, 550–558.

[5] Petra Badke-Schaub, Gabriela Goldschmidt, and Martijn Meijer. 2010. How does cognitive conflict in design teams support the development of creative ideas? *Creativity and Innovation Management* 19, 2 (2010), 119–133. https://doi.org/10.1111/j.1467-8691.2010.00553.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8691.2010.00553.x

[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81.

[7] Sumit Basu, Danyel Fisher, Steven Drucker, and Hao Lu. 2010. Assisting users with clustering tasks by combining metric learning and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 394–400.

[8] Frédéric Bevilacqua, Norbert Schnell, Nicolas Rasamimanana, Julien Bloit, Emmanuel Flety, Baptiste Caramiaux, Jules Françoise, and Eric Boyer. 2013. De-MO: Designing action-sound relationships with the MO interfaces. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 2907–2910.

[9] Parzival Borlinghaus and Stephan Huber. 2021. Comparing apples and oranges: Human and computer clustered affinity diagrams under the microscope. In *26th International Conference on Intelligent User Interfaces (IUI'21)*. Association for Computing Machinery, New York, NY, USA, 413–422. https://doi.org/10.1145/3397481.3450674

[10] Michael E. Bratman. 1992. Shared cooperative activity. *The Philosophical Review* 101, 2 (1992), 327–341. https://doi.org/10.2307/2185537

[11] Anni Coden, Marina Danilevsky, Daniel Gruhl, Linda Kato, and Meena Nagarajan. 2017. A method to accelerate human in the loop clustering. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 237–245.

[12] Nigel Cross. 2004. Expertise in design: An overview. *Design Studies* 25, 5 (2004), 427–441.

[13] Raymond H. Cuijpers and Marco A. M. H. Knops. 2015. Motions of robots matter! The social effects of idle and meaningful motions. In *International Conference on Social Robotics*. Springer, 174–183.

[14] Nicholas Davis, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An enactive model of creativity for computational collaboration and co-creation. In *Creativity in the Digital Age*. Springer, 109–133.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology* 53, 3 (1987), 497.

[17] Kees Dorst. 2011. The core of 'design thinking' and its application. *Design Studies* 32, 6 (2011), 521–532.

[18] Steven M. Drucker, Danyel Fisher, and Sumit Basu. 2011. Helping users sort faster with adaptive machine learning recommendations. In *IFIP Conference on Human-Computer Interaction*. Springer, 187–203.

[19] Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. 2017. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 382–391.

[20] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, Vol. 96. 226–231.

[21] Patrick Faion, Pablo Prietz, Petrov Aleksandar, Rohit Suri, Roman Roibu, and Joseph Moster. 2022. Pupil-Apriltags: Python Bindings for the apriltags3 Library. https://github.com/pupil-labs/apriltags

[22] Florian Geyer, Ulrike Pfeil, Jochen Budzinski, Anita Höchtl, and Harald Reiterer. 2011. AffinityTable — A hybrid surface for supporting affinity diagramming. In *Human-Computer Interaction—INTERACT 2011 (Lecture Notes in Computer Science)*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer, Berlin, 477–484. https://doi.org/10.1007/978-3-642-23765-2_33

[23] Gabriela Goldschmidt. 1991. The dialectics of sketching. *Creativity Research Journal* 4, 2 (Jan. 1991), 123–143. https://doi.org/10.1080/10400419109534381

[24] Kazjon Grace and Mary Lou Maher. 2016. Surprise-triggered reformulation of design goals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[25] Jesse Gray, Guy Hoffman, Sigurdur Orn Adalgeirsson, Matt Berlin, and Cynthia Breazeal. 2010. Expressive, interactive robots: Tools, techniques, and insights based on collaborations. In *HRI 2010 Workshop: What Do Collaborations with the Arts Have to Say About HRI*. 21–28.

[26] Christian Guckelsberger, Christophe Salge, and Simon Colton. 2017. Addressing the "why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 128–135.

[27] Gunnar Harboe. 2013. Understanding and augmenting a paper arrangement-based method. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp'13 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 343–348. https://doi.org/10.1145/2494091.2501087

[28] Gunnar Harboe and Elaine M. Huang. 2015. Real-world affinity diagramming practices: Bridging the paper-digital gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2702123.2702561

[29] Gunnar Harboe, Jonas Minke, Ioana Ilea, and Elaine M. Huang. 2012. Computer support for collaborative data analysis: augmenting paper affinity diagrams. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW'12)*. Association for Computing Machinery, New York, NY, USA, 1179–1182. https://doi.org/10.1145/2145204.2145379

[30] Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. Sage Publications Sage CA: Los Angeles, CA, 904–908.

[31] 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.

[32] Julie Heiser, Barbara Tversky, and Mia Silverman. 2004. Sketches for and from collaboration. *Visual and Spatial Reasoning in Design III* 3 (2004), 69–78.

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

[34] Guy Hoffman, Rony Kubat, and Cynthia Breazeal. 2008. A hybrid control system for puppeteering a live robotic stage actor. In *RO-MAN 2008 — The 17th IEEE International Symposium on Robot and Human Interactive Communication*. 354–359. https://doi.org/10.1109/ROMAN.2008.4600691

[35] Eva Hornecker and Jacob Buur. 2006. Getting a grip on tangible interaction: A framework on physical space and social interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 437–446.

[36] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems*. 57–64.

[37] Matthew Huggins, Sharifa Alghowinem, Sooyeon Jeong, Pedro Colon-Hernandez, Cynthia Breazeal, and Hae Won Park. 2021. Practical guidelines for intent recognition: BERT with minimal training data evaluated in real-world HRI application. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 341–350.

[38] Tim Ingold. 2010. The textility of making. *Cambridge Journal of Economics* 34, 1 (Jan. 2010), 91–102. https://doi.org/10.1093/cje/bep042

[39] Hiroshi Ishii. 2008. Tangible bits: Beyond pixels. In *Proceedings of the 2nd International Conference on Tangible and Embedded Interaction*. xv–xxv.

[40] Robert J. K. Jacob, Audrey Girouard, Leanne M. Hirshfield, Michael S. Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. 2008. Reality-based interaction: A framework for post-WIMP interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. Association for Computing Machinery, New York, NY, USA, 201–210. https://doi.org/10.1145/1357054.1357089

[41] Peter H. Kahn, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Solace Shen, Jolina H. Ruckert, and Heather E.Gary. 2016. Human creativity can be facilitated through interacting with a social robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 173–180.

[42] Pegah Karimi, Kazjon Grace, Nicholas Davis, and Mary Lou Maher. 2018. Creative sketching apprentice: Supporting conceptual shifts in sketch ideation. In *International Conference on Design Computing and Cognition*. Springer, 721–738.

[43] Hee-Su Kim and Sung-Bae Cho. 2000. Application of interactive genetic algorithm to fashion design. *Engineering Applications of Artificial Intelligence* 13, 6 (2000), 635–644.

[44] S. R. Klemmer, Mark Newman, F. Farrell, Raecine Meza, and James A. Landay. 2000. A tangible difference: Participatory design studies informing a designers' outpost. In *CSCW 2000 Workshop on Shared Environments to Support Face-to-Face Collaboration*. Citeseer.

[45] Scott R. Klemmer, Björn Hartmann, and Leila Takayama. 2006. How bodies matter: Five themes for interaction design. In *Proceedings of the 6th Conference on Designing Interactive Systems*. 140–149.

[46] Scott R. Klemmer, Mark W. Newman, Ryan Farrell, Mark Bilezikjian, and James A. Landay. 2001. The Designers' Outpost: A tangible interface for collaborative web site. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*. 1–10.

[47] Janin Koch and Antti Oulasvirta. 2016. Computational layout perception using gestalt laws. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1423–1429.

[48] John Kruschke. 2015. Doing Bayesian data analysis (Second Edition). Boston: Academic Press.

[49] John K. Kruschke. 2018. Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science* 1, 2 (2018), 270–280.

[50] Matthew V. Law, JiHyun Jeong, Amritansh Kwatra, Malte F. Jung, and Guy Hoffman. 2019. Negotiating the creative space in human-robot collaborative design. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS'19)*. Association for Computing Machinery, New York, NY, USA, 645–657. https://doi.org/10.1145/3322276.3322343

[51] Marianne Aubin Le Quere, Maria Antoniak, Tegan Wilson, Alexa VanHattum, Grin Berlstein, Elizabeth Ricci, Andrea Cuadra, Sachi Angle, and Sharifa Sultana. 2020. COVID-19 graduate students in computing at Cornell survey results. *Graduate Students for Gender Inclusion in Computing* (2020). https://drive.google.com/file/d/1hC5CIdvHRVXJFf4uiikkDbGQHGmpfnsM

[52] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. 2006. Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of Communication* 56, 4 (2006), 754–772.

[53] Jingxian Liao and Hao-Chuan Wang. 2019. Gestures as intrinsic creativity support: Understanding the usage and function of hand gestures in computer-mediated group brainstorming. *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–16.

[54] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It is your turn: Collaborative ideation with a co-creative robot through sketch. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376258

[55] Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14–18, 2015, Proceedings, Part II 15*. Springer, 231–248.

[56] Michael Nunes, Saul Greenberg, and Carman Neustaedter. 2008. Sharing digital photographs in the home through physical mementos, souvenirs, and keepsakes. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems*. 250–260.

[57] Paul Paulus. 2000. Groups, teams, and creativity: The creative potential of idea-generating groups. *Applied Psychology* 49, 2 (2000), 237–262. https://doi.org/10.1111/1464-0597.00013 eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1464-0597.00013

[58] Pallets Projects. 2010. Flask Web Server Documentation. Retrieved from https://flask.palletsprojects.com/

[59] Juan C. Quiroz, Sushil J. Louis, Amit Banerjee, and Sergiu M. Dascalu. 2009. Towards creative design using collaborative interactive genetic algorithms. In *2009 IEEE Congress on Evolutionary Computation*. IEEE, 1849–1856.

[60] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[61] Donald A. Schön. 1992. Designing as reflective conversation with the materials of a design situation. *Knowledge-based Systems* 5, 1 (1992), 3–14. Publisher: Elsevier.

[62] Raymond Scupin. 2008. The KJ method: A technique for analyzing data derived from Japanese ethnology. *Human Organization* 56, 2 (Jan. 2008), 233–237. https://doi.org/10.17730/humo.56.2.x335923511444655

[63] Herbert A. Simon. 2019. *The Sciences of the Artificial*. MIT Press.

[64] Tony L. Simons and Randall S. Peterson. 2000. Task conflict and relationship conflict in top management teams: The pivotal role of intragroup trust. *Journal of Applied Psychology* 85, 1 (2000), 102.

[65] Hariharan Subramonyam, Steven M. Drucker, and Eytan Adar. 2019. Affinity lens: Data-assisted affinity diagramming with augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300628

[66] John C. Tang. 1991. Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies* 34, 2 (1991), 143–160.

[67] Barbara Tversky and Bridgette Martin Hard. 2009. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition* 110, 1 (Jan. 2009), 124–129. https://doi.org/10.1016/j.cognition.2008.10.008

[68] Barbara Tversky and Angela Kessell. 2014. Thinking in action. *Pragmatics & Cognition* 22, 2 (Jan. 2014), 206–223. https://doi.org/10.1075/pc.22.2.03tve

[69] UFactory. 2016. PyUArm (uArm Metal). Retrieved from https://github.com/uArm-Developer/pyuarm

[70] Daniella G. Varela and LaVonne C. Fedynich. 2021. Teaching from a social distance: Teacher experiences in the age of COVID-19. *Research in Higher Education Journal* 39 (Jan. 2021). https://eric.ed.gov/?id=EJ1293887

[71] Willemien Visser. 2009. Function and form of gestures in a collaborative design meeting. In *International Gesture Workshop*. Springer, 61–72.

[72] Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. 2011. Integrating document clustering and multidocument summarization. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5, 3 (2011), 1–26.

[73] John Wang and Edwin Olson. 2016. AprilTag 2: Efficient and robust fiducial detection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4193–4198.

[74] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.

Received 8 November 2022; revised 30 September 2023; accepted 28 December 2023