AN ABSTRACT OF THE DISSERTATION OF

Karina A. Roundtree for the degree of Doctor of Philosophy in Mechanical Engineering presented on August 21, 2020.

Title: Achieving Transparency in Human-Collective Systems

Abstract approved: _____

H. Onan Demirel Julie A. Adams

Collective robotic systems are biologically-inspired and exhibit behaviors found in spatial swarms (e.g., fish), colonies (e.g., ants), or a combination of both (e.g., bees). Collective robotic system popularity continues to increase due to their apparent global intelligence and emergent behaviors. Many applications can benefit from the incorporation of collectives, including environmental monitoring, disaster response missions, and infrastructure support. Human-collective system designers continue to debate how best to achieve transparency in human-collective systems in order to attain meaningful and insightful information exchanges between the operator and collective, enable positive operator influence on collectives, and improve the human-collective's performance.

Few human-collective evaluations have been conducted, many of which have only assessed how embedding transparency into one system design element (e.g., models, visualizations, or control mechanisms) may impact human-collective behaviors, such as the human-collective performance. This dissertation developed a transparency definition for collective systems that was leveraged to assess how to achieve transparency

in a single human-collective system. Multiple models and visualizations were evaluated for a sequential best-of-n decision-making task with four collectives. Transparency was evaluated with respect to how the model and visualization impacted human operators who possess different capabilities, operator comprehension, system usability, and human-collective performance. Transparency design guidance was created in order to aid the design of future human-collective systems. One set of guidelines were inspired from the results and discussions of the single human-collective analyses and another set were based on a review of the biological literature.

This dissertation can be used to aid designers achieve transparency in human-collective systems. The primary contributions are:

1. A transparency definition for human-collective systems that describes the process of identifying what factors affect and are influenced by transparency, why those factors are important, and how to design a system to achieve transparency.

2. An expansive set of metrics that successfully evaluated how transparency influenced operators with different individual capabilities, operator comprehension, system usability, and human-collective performance.

3. The recommendation that system transparency quantification requires evaluating the transparency embedded into the various system design elements in order to determine how they interact with one another and influence the human-collective interactions and performance.

4. Design guidance recommendations with respect to models, visualizations, and control mechanisms in order to inform designers how transparency can be achieved for human-collective systems.

Achieving Transparency in Human-Collective Systems

by

Karina A. Roundtree

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 21, 2020
Commencement June 2021

Doctor of Philosophy dissertation of Karina A. Roundtree presented on August 21, 2020.

APPROVED:

Co-Major Professor, representing Mechanical Engineering

Co-Major Professor, representing Mechanical Engineering

Head of the School of Mechanical, Industrial, and Manufacturing Engineering

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Karina A. Roundtree, Author

ACKNOWLEDGEMENTS

I completed this dissertation as a graduate research assistant supported by the US Office of Naval Research Awards N000141613025 and N000141812831. While completing this work, I received technical assistance from Dr. Michael Goodrich, Dr. Jason R. Cody, Dr. Jamison Heard, Dr. Gina Olson, Dr. Gilberto Marcon dos Santos, Jennifer Leaf, Dagim Gebretsadik, Lorenzo Bermillo, Golden Rockefeller, Rene Martinez, Alan Sanchez, Anita Ruangrotsakun, and Gabriela Magana. Thank you all for your support.

The most important person I must thank for helping me through my PhD journey is my partner, Dagim Gebretsadik. My success throughout this process is attributed to you. Thank you for being there every single step of the way, being my biggest fan, for always believing in me, and for your support, love, guidance, and help (especially when it came to coding). I love you dearly. This PhD belongs to both of us!

The second set of people I want to thank are my advisors, Dr. Onan Demirel and Dr. Julie A. Adams. Thank you Dr. Demirel for your support, enthusiasm, and for always sparking my creativity. I was able to explore and learn about many domains thanks to your guidance. Thank you for always believing in me. Julie, there are no words that can express how thankful and blessed I am to have had the opportunity to work with you. The amount of professional and technical growth I have gained from you is tremendous. Thank you for giving me opportunities to grow and for constantly challenging me out of my comfort zone.

I want to thank all the professors who aided me throughout this process. Thank you Dr. Michael Goodrich for being an amazing collaborator on the ONR grant and for your insights, guidance, and support on my research. Dr. Nancy Squires thank you

always believing, supporting, and inspiring me. You are truly missed. Thank you Dr. Irem Tumer for helping find a PhD project I fell in love with. Thank you Dr. Ken Funk for entertaining my questions in class and supporting me. I learned so much from you. I also want to thank my committee, for their time, effort, and insights.

There is a long list of OSU students I want to thank from the Mechanical Engineering, Industrial Engineering, and Robotics program that supported me. Your friendship, kindness, and support helped me during this PhD journey. I want to thank my friends and sorority sisters, especially Daniella Roquett, for always cheering me on from afar and supporting my family and I through some hard personal times. The last and most important people I must thank is my family. Thank you for your love, patience, support, encouragement, and for always believing in me. It has been a long journey. We finally made it, te quiero.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF APPENDICES

LIST OF APPENDIX TABLES

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF ABBREVIATIONS

A Assessing.

AT all timings associated with SA probe questions.

B 15 seconds before asking a SA probe question.

C committed.

D during response to a SA probe question.

Dec Diff decision difficulty.

DOF degree of freedom.

EP Engagement Prompts.

F feedback.

FB Feedback.

GAICE area corresponding to the 800 individual collective entities for global clutter percentage calculation.

GCIW area corresponding to the number of collective information pop-up windows for global clutter percentage calculation.

GHA area corresponding to the four collective hubs for global clutter percentage calculation.

GHTA area corresponding to the number of highlighted targets for global clutter percentage calculation.

GTA area corresponding to the number of not highlighted targets for global clutter percentage calculation.

GTIW area corresponding to the number of target information pop-up windows for global clutter percentage calculation.

H hypotheses.

Hz Hertz.

IA Individual Agents.

ICA area corresponding to the static interface components for global clutter percentage calculation.

LAICE area corresponding to the number of individual collective entities for local clutter percentage calculation.

LCIW area corresponding to the number of collective information pop-up windows for local clutter percentage calculation.

LHA area corresponding to the number of collective hubs for local clutter percentage calculation.

LHTA area corresponding to the number of highlighted targets for local clutter percentage calculation.

LTA area corresponding to the number of not highlighted targets for local clutter percentage calculation.

LTIW area corresponding to the number of target information pop-up windows for local clutter percentage calculation.

max maximum.

min minimum.

NASA-TLX NASA Task Load Index.

P Providing.

PM Planning Mechanisms.

S Status.

SA situation awareness.

SART 3-D Situational Awareness Rating Technique.

SAT Situation awareness-based Agent Transparency.

SD standard deviation.

U uncommitted.

W while being asked a SA probe question.

X executing.

Chapter 1: Introduction

The design of complex systems impacts how efficiently and effectively humans and machines accomplish tasks in particular environments. Aspects associated with the human, the machine, and the interactions that occur between the entities, must be considered in order to maximize desired outcomes, such as high performance. Transparency, the principle of providing easily exchangeable information, such as shared awareness, intent, and reasoning processes, in order to enhance comprehension, is necessary to provide meaningful and insightful information exchanges between the system and the human [2]. Effective human-machine team interactions, such as communication, cooperation, and ultimately the team's performance, are influenced by transparency. This dissertation developed and evaluated methods of achieving transparency for human-collective teaming systems, evaluated transparency metrics, and developed design guidelines for practical future use scenarios.

Collective robot systems are biologically inspired and exhibit behaviors found in spatial swarms (e.g., [3]), colonies (e.g., [4]), or a combination of both (e.g., [5]). A honeybee colony searching for a new hive location is an example of collective behavior. Initially, a subset of the colony population leaves the hive in search of a new hive for a daughter colony [5]. The subset of honeybees fly a short distance before coalescing, often on a tree branch, where they wait while a set of scout honeybees search the area for a new hive location. During the initial flight, the honeybees exhibit spatial swarm behaviors typically found in flocks of birds [6] or schools of fish [7], where each honeybee maintains a particular distance from their neighbor in order to avoid collisions and fol-

low their neighbor in a particular direction. Scout honeybees explore the surrounding area for possible hive alternatives, each evaluating the alternatives with respect to the ideal hive criteria. The scouts return to the waiting swarm to begin a selection process (i.e., colony behavior) entailing debate, agreement on a best hive location (i.e., best-of-n [8]), and building consensus. After completing their consensus decision-making process, all the honeybees travel to the new hive location, transitioning from colony based behaviors back to spatial swarm behaviors.

Collective robot systems are technologically simple. System attributes, such as collective intelligence and emergent behaviors, make these systems advantageous for task completion, because they are: (1) scalable (i.e., can change in size) [3], (2) resilient to failures (i.e., responsibilities can be redistributed to other collective entities) [9], and (3) flexible in varying environments [10, 11], as well as the type of robotic entities used (i.e., heterogeneous members). Many applications can benefit from the incorporation of collective robot systems, including environmental monitoring, disaster response missions, infrastructure support, and protection [3]. This dissertation focused on human-collective systems that include spatial swarms, colonies, and hub-based collectives. A hub is a centralized point, similar to a honeybee hive, where the collectives' individual entities gather to exchange information, receive tasks, refuel, and undergo repairs.

Transparency results from traditional human-machine domains [2] informed design requirements to be integrated into the human-collective system design. The resulting system was evaluated in order to assess the impacts on operators with different individual capabilities, their comprehension, the system usability, and the human-collective's performance. Understanding how the transparency embedded in different system design elements influence interactions between the operators and collectives was used to inform transparency focused design guidelines for human-collective systems.

The means of providing transparency for human supervisors when monitoring and tasking distributed hub-based collectives is challenging. The quantity and quality of insightful information provided to the operators will impact the perception of the system's state, which includes the collective's overall mission completion status, resource allocation, as well as what the system is currently doing, what it plans to do, and any other future predictive information. The overall understanding and transparency of the system can be constrained when too little information is available, while too much information may overload the operators. Determining the correct level of information necessary is critical to designing optimal transparent human-collective systems.

The ability to provide transparency will become more challenging as the complexity of the human-collective system increases. Understanding how the human, system, and environment interact with one another identified crucial aspects to improve the efficacy of human-collective interactions. The identified aspects have the potential to improve transparency and desired system outcomes, such as enabling operators with different individual capabilities to perform relatively the same, as well as promoting operator comprehension, system usability, and optimal human-collective performance.

This dissertation contributed novel evaluated methods to achieve human-collective transparency, as well as developed design requirements that better support future human-collective systems. Chapter 2 summarizes the existing transparency research and metrics used to assess transparency from various human-machine domains, as well as provides background information related to biomimicry that was used for deriving design guidance. Chapter 3 focuses on the experimental design of the single operator-collective evaluations. The evaluation analyses hypotheses, results, and discussion for each research question are presented in Chapter 4. Transparency design guidelines, presented in Chapter 5, were created from the single operator-collective evaluation results and a

review of the biomimicry literature. The conclusions, dissertation contributions, and future work are provided in Chapter 6. The questionnaires used prior, during, and at completion of the single operator-collective evaluations, as well as additional results not presented in Chapter 4, are provided in the Appendices.

Chapter 2: Literature Review

The development of highly complex systems are emerging at a rapid rate in various domains, such as transportation and robotics. These more sophisticated systems challenge designers to consider what factors influence the human operator, the machine, and the environment. Understanding the interactions that occur between these entities, shown in Figure 2.1, are necessary in order to achieve high efficiency, productivity, and safety.

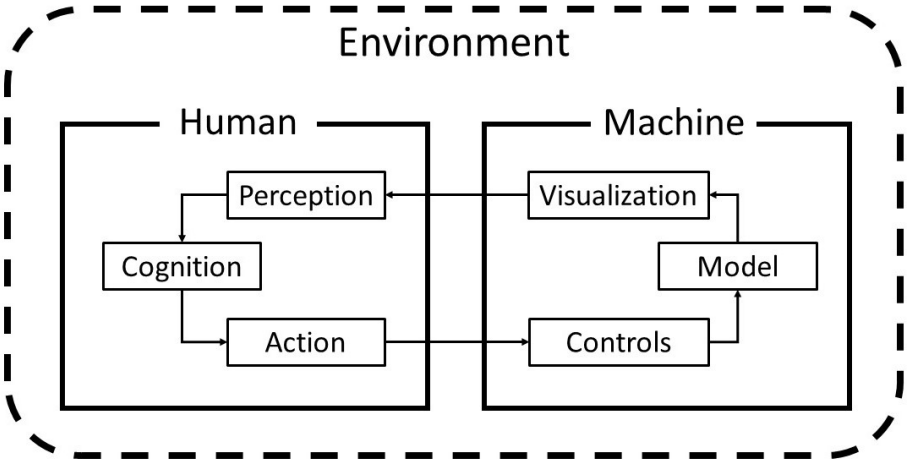


Figure 2.1: Human-machine system operating in an environment [1].

Examples of unique behaviors and characteristics associated with biological spatial swarms and colonies are discussed, which were used as inspiration for the development of design guidance for human-collective systems in Chapter 5.2. Definitions of the robotic spatial swarm, colony, and collective systems are provided, as well as the current visualization and interactions transparency research relative to a particular robotic

system. The collective literature that exemplifies both spatial swarm and colony behaviors embedded in one collective system is limited. Only one collective interactions transparency evaluation has been conducted om the existing literature. This dissertation expands on that literature and is discussed further in Chapter 3. The literature includes evaluations where an operator served as a supervisor [12] aiding simulated spatial swarms, colonies, or collectives. Definitions of transparency are provided along with direct and indirect factors that are either influenced by transparency or affect transparency. Three primary design methods (provide, design, and train), that can be used to promote or embed transparency into systems, are discussed.

2.0.1 Spatial Swarms

Biological spatial swarms are comprised of a large number (>50) of simplified members that exhibit intelligent and emergent behaviors as a unit. Behaviors unique to biological spatial swarms are provided in order to understand of what characteristics may be important to incorporate into robotic spatial swarms. Understanding how spatial swarm individual entities communicate and interact with one another to influence other individual entity and global spatial swarm state changes is necessary to ground human-spatial swarm system design.

2.0.1.1 Biological Spatial Swarm Behaviors

Six behaviors were identified from the spatial swarm literature that contribute to characteristics discussed in the spatial swarm definition (Chapter 2.0.1.2) and helped inspire design guidance for future human-collective systems (Chapter 5.2). The first behavior

is cohesion, which is the degree of connectedness in a group [13]. The most common benefit of cohesion is increased safety. Honeybees, for example, take off and fly together as a cohesive group towards their new nest site in order to mitigate risks from their environment [5]. Higher aggregation cohesion helps lower the number of isolated individuals or small groups of birds from attacks, such as those from falcons [6]. Fish that align with each other, rather than away from a predator (e.g., shark), can avoid collisions with one another and contribute to creating a confusion effect, whereby the predator is unable to focus on any single individual [14]. Tight aggregation cohesion can result from selfish individuals using others to their benefit, such as the described cover from predators [15]. Cohesion can be maintained in a group by having a fixed number of individuals interacting with one another. The shape and density can fluctuate, all while maintaining some degree of cohesion [6], allowing for complex geometry, such as parabolic formations used for cooperative hunting in tuna [16]. Ideal reshuffle rates, where individuals in a spatial swarm change positions amongst each other, are necessary to maintain long-range cohesive order [17]. Members who reshuffle too quickly can cause detrimental effects to the spatial swarm.

Individual roles may persist or change, as determined by various characteristics. Physiological characteristics can influence particular behaviors, such as fish body length determining group size distribution [18] and nutritional state determining where individuals are placed within a group [14]. Well-fed fish, for example, will move towards the center of the group, while hungry individuals move towards the outside [19]. Environmental characteristics can also influence behaviors, such as birds located on the border of a spatial swarm tend to exchange positions with neighbors less than those internal to the spatial swarm [17]. Individuals inherit some roles due to the collective's need. One bird can act as a sentinel, for example, while the others feed, which can help

reduce the probability of an unnoticed predator attack [20].

Communication among individuals is typically limited in biological spatial swarms, and the information provided by the members can be presented at the collective or subgroup level. The number of neighbors individuals' communicate with varies depending on the species. Starlings communicate with six to seven neighbors [6], while shoaling fish communicate with three to five [21]. Differences in the number of neighbors may be caused by various characteristics. Fish, for example, may visually occlude others' ability to perceive neighbors [22] due to their size or the density of the aggregation. Interacting with few neighbors reduces the noisiness of the information, at the expense of the information communication range [6]. Birds implement strategies, such as reshuffling as a way to change the neighbors with whom they interact over time [6].

Characteristics, such as how informed and experienced individuals are, their proximity to others, and physiology, can determine which individual spatial swarm entities become leaders in spatial swarms. Streakers are informed scout honeybees that provide flight direction information towards a nest site to the other uniformed honeybees by flying at the top of the swarm where they can easily be seen against the sun [5]. Scout honeybees partake in the consensus decision-making process to choose a new nest site; therefore, scout honeybees that have visited the chosen site know where it is located and can lead others to it. Experienced fish, that have been trained to do a particular task, can influence a naïve group of fish to do the same task [23], suggesting that individuals who are experienced can become leaders. The information provided by close proximity neighbors has been weighted as more important than the information provided by neighbors that are further away [24], since perception of others decreases with increased distance. Better-nourished individuals who forage with poorer-nourished individuals stop the foraging process after they no longer gain any benefit from foraging

(i.e., the better-nourished individual is full) [25], which demonstrates how physiological characteristics can cause some individuals to lead decision-making.

Leadership for many species is transient and can change due to various circumstances, such as large population sizes, long distances that must be travelled, predictable resources, and navigational habits, as is demonstrated in dolphins [26]. Leaders do not need to consistently lead others throughout the entire task duration time, which has been observed by dolphin leaders who typically spend less than 20% of their time leading. Having a small number of leaders may help reduce discrepancies that can arise if too many leaders are making decisions, which improves decision-making time and saves energy during travel. Individuals can also become leaders depending of the time of the year, which has been observed in heifers [27]. Transient leadership has many advantages and must be considered in the design guidelines.

2.0.1.2 Definition of Robotic Spatial Swarms

Spatial swarm robots are biologically inspired by self-organized social animals [3], such as flocks of birds [6] and schools of fish [7]. Spatial swarms are comprised of a large number ($>$50) of simplified individual entities that exhibit intelligent, emergent behaviors as a unit, and respond to locally available information among the individual swarm entities to achieve an objective [28–30]. The spatial swarm's global intelligence and emergent behaviors make it scalable, resilient, and flexible [10, 11]. *Scalability* refers to the spatial swarm's ability to perform well, regardless of its size [3]. Individual entities may suffer failures; however, the spatial swarm is *resilient*, as responsibilities can be redistributed to others in order to achieve the task [9]. The spatial swarm's ability to adapt to varying environments and tasks represents *flexibility* [9].

Robotic spatial swarms often rely on distributed, localized, often implicit communication, as well as basic rules of repulsion, attraction, and orientation that enable individual entities to position themselves relative to neighboring entities [31]. Couzin et al.'s [31] model states that individual entities in the zone of repulsion attempt to maintain a minimum distance from their neighbors, striving to avoid collisions. The zone of orientation causes individual entities to align themselves with neighbors who are in close proximity. Entities that are far from their neighbors will move closer, as a factor of the zone of attraction. Relative motion among spatial swarms, such as those observed in fish, enrich visually driven communication [32]. Methods to communicate information across a spatial swarm include, salient movements warning individual entities that a predator is in proximity [33] or rapid changes in acceleration, such as a streaking honeybee guiding a spatial swarm in a particular direction [5].

2.0.1.3 Visualization Transparency for Human-Spatial Swarm Systems

The spatial swarm transparency literature has typically implemented traditional visualizations (i.e., showing the position of every individual spatial swarm entity) has focused on assessing operator understanding of spatial swarm behaviors and human-spatial swarm performance. A variety of individual entity features have been visualized for operators, including current and predicted future position [34] and heading direction [35], health (i.e., speed, strength, capability, and dispersion [36]), and status. The most commonly used visual icon for an individual entity has been a circle, where directional information was observed either as the entity moves across a 2-D space, or the cir-

cle incorporates a line pointing in the heading direction. Existing visualization design guidance was derived from subject matter experts [36] and the Gestalt principles [37]. The use of multimodal cues (i.e., visual spatial swarm state color coding and written messages, spoken messages, and vibrations) aided operators in the identification and response of signals during a reconnaissance mission (99.9% accuracy), and resulted in shorter response times, and lower workload. Sharing spatial swarm information via multimodal cues may alleviate an operator's high visual loads when managing multiple tasks on various displays by increasing situational awareness [36] and promoting better transparency. Operators' using a visualization incorporating Gestalt-based design principles perceived and approximated optimal spatial swarm performance faster than operators using visualizations containing only individual entity position information [37]. Increased visualization transparency enabled operators to learn when to approximate optimal input timing.

Information latency, which can occur due to communication bandwidth limitations, and neglect benevolence, the time allowed for a spatial swarm to stabilize before issuing new operator commands, on operator understanding of future spatial swarm behaviors are important considerations [34]. Latency affected the operators' ability to control a spatial swarm, but providing additional transparency via a predictive visualization, which showed each individual entity's predicted location 20 seconds into the future, mitigated these effects. Operators using the predictive visualization with latency performed as well as operators who experienced no latency. Transparency of human-spatial swarm systems can be improved by implementing predictive visualizations of the spatial swarm and its entities by allowing the operator more time to think about their future actions. Operators will be able to balance span, the number of individual entities they can interact with, and persistence, the duration of the interactions

with individual entities, by using visualizations that provide heading information [35]. Aspects, such as the presence or absence of Couzin et al.'s [31] communication model states, visualized individual entities' velocity, and individual operator characteristics, such as gender, have impacted the identification of spatial swarming behavior [38]. Understanding the influence of factors is necessary to promote perception and comprehension of spatial swarm behavior to inform future operator actions.

Abstract spatial swarm visualizations have been proposed to improve operator understanding and influence positive spatial swarm behavior. Radial visualizations, using the three level Situation awareness-based Agent Transparency (SAT) framework [39] and heuristic evaluations analyzing the application of spatial swarm metrics on visualizations [13], as well as glyphs [40], bounding ellipses [41], convex hulls, and directed arrows [42] have been assessed. Operators using a glyph were able to acquire information regarding the spatial swarm's power levels, task type, and the number of individual entities, via one icon [40]. Additional information about particular system features was accessible via pop-up windows. Designers of abstract spatial swarm visualizations can ensure transparency by providing redundant information via the spatial swarm icon and using supplementary information windows.

Conflicting results were found for evaluations assessing whether traditional or abstract visualizations aided operators better during different tasks. Abstract visualizations during a go-to and avoid task in the presence or absence of obstacles [42] performed worse than the traditional visualizations, while abstract visualizations performed as well or better when perceiving biological spatial swarm structure [43] and under variable bandwidth conditions [41]. Further analysis is needed in order to determine which spatial swarm visualization will promote better transparency for a common task by investigating how transparency factors influence human-spatial swarm behaviors.

2.0.1.4 Influence of Transparency on Human-Spatial Swarm Interactions

Many of the existing transparency evaluations investigated how control mechanisms influenced human-spatial swarm interactions and behavior (e.g., [44, 45]). Two mechanisms were used to control a spatial swarm foraging in simple and complex environments [44, 46]. Selection, influenced a selected subgroup, and beacon, exerted influence on entities within a set range. The highest performance occurred when fully autonomous spatial swarms (i.e., no operator influence) foraged in simple environments, while selection was optimal in complex environments [44]. Selection generally outperformed beacon; however, as the spatial swarm size increased, beacon became more advantageous by requiring less operator influence [46]. Improvements must be considered in order to reduce the learning curve of using beacon and improve its effectiveness (i.e., learning where to strategically place beacons).

Leader, predator, and mediator control mechanisms were assessed, with regard to spatial swarm manageability and performance [45]. Leaders attracted entities towards themselves, predators repelled entities away, and mediators allowed the operator to mold and adapt the spatial swarm. Operators experienced different workload levels and implemented different control strategies depending on the control mechanism. Workload increased when using leaders, decreased with predators, and remained relatively stable with mediators. Operators using leaders gathered all of the spatial swarm entities together and guided them in a particular direction. Spatial subswarms emerged and were pushed in different directions when the operators used predators. Mediators were strategically placed in the environment, which resulted in lower workload, suggesting that this control mechanism may be easier to use.

The quantity and quality of operator influence was investigated to identify when the influence begins to have a detrimental effect on human-spatial swarm performance [47]. Operators moved a spatial swarm around in various environments at two levels of autonomy using an autonomous dispersion algorithm (high autonomy) and user-defined goto points (low autonomy). Operator influence was required in complex environments containing numerous obstacles and small passageways; however, too much control never allowed the autonomy to operate, resulting in a performance decline. Two operator interaction strategies emerged: (1) allow the autonomous algorithm to control spatial swarm movement or (2) manually break the spatial swarm into subgroups and guide them to explore different areas of the map.

Two evaluations assessed the influence of visualizations on human-spatial swarm interactions. Four methods of displaying the spatial swarm's state were assessed based on the operator's ability to predict the spatial swarm's future state [41]. The full information display showed the position and heading of each individual entity, the centroid/ellipse showed a bounding ellipse at the center of the spatial swarm, the minimum volume enclosing ellipse showed leaders at the edge of the spatial swarm, and random condition clustering showed leaders evenly spaced throughout the spatial swarm. The full information and centroid/ellipse displays enabled the most accurate predictions when estimating spatial qualities, with a preference for the bounding ellipse in low bandwidth situations. The leader-based strategies may be more advantageous for tasks that have a goal, such as the best-of-n decision-making task, which is a selection process entailing debate and building consensus on the best hive location out of n options [8]. A metacognition model that enabled individual entities to monitor changes in the spatial swarm's state and a visualization that communicated spatial swarm status during a convoy mission were assessed when information was provided in different

modalities (e.g., spatial, audio, and tactile cues) in order to increase situational awareness of surroundings and improve visual attention [36]. The task required monitoring the spatial swarm and responding to display signals, while performing a robotic planning task. The visualization enabled 99.9% accuracy of signal detection and recognition.

2.0.2 Colonies

The unique biological colony behaviors are provided in order to understand what characteristics may be important when designing systems to provide transparency for robotic colonies. Determining how colony members communicate and interact with one another to influence other individuals and the global colony state is necessary to ground human-colony system design.

2.0.2.1 Biological Colony Behaviors

Eight behaviors, five previously mentioned behaviors from Chapter 2.0.1.1 and four new behaviors, were identified from the colony literature that contribute to characteristics discussed in the colony definition (Chapter 2.0.2.2) and helped inspire design guidance for future human-collective systems (Chapter 5.2). The previous five behaviors and characteristics specific to colonies are discussed first, followed by the new behaviors. The benefits of cohesion in spatial swarms and the desire to maintain cohesion to increase safety also apply to colonies. Honeybee colonies, for example, will aggregate into tight, well-insulated clusters in order to survive winter temperatures [5].

Roles are often more clearly defined in colonies than spatial swarms. Particular members, such as worker honeybees, perform a variety of roles, which vary with age,

include cleaning cells to feeding larvae, building combs, ventilating the hive, guarding the entrance, and foraging [5]. SiFmilar roles are performed by ants that find food and bring it back to the nest, build and repair the nest, as well as feed and groom the larvae [48]. These colony members often proactively make inspections searching for things to do based on the colony's needs [5]. Environmental and physiological characteristics can also influence particular role behaviors. Conditions inside a honeybee colony's hive (congestion of the adult honeybees, numerous immature honeybees, and expanding food reserves) and outside the hive (plentiful pollen during the spring), for example, have been correlated with worker honeybees starting the process of queen rearing [5]. The changing environment also influences ants that react by performing other tasks [48]. As honeybees and ants age, their roles change. Nest-site scouts are elderly honeybees that served previously as common foragers [5]. A general pattern of role change in ants starts with younger workers staying inside the nest, working on brood care and nest construction, and then moving to work outside the nest, where they forage for food when they are older [48].

Communication for colonies can occur inside a hive or nest, outside using strategies similar to spatial swarm communication, or can be embedded into the environment. Communication among colony members is also limited, like the spatial swarms, and the information provided by the members can be presented at the collective or subgroup level. Honeybees can only observe and react to the actions of their immediate neighbors; hence, honeybees operate without global knowledge of the information that percolates among other fellow honeybees [5]. Local sampling, performed in parallel by large numbers of individuals, allows the colony to accurately tune its average response to environmental changes [14]. The group-level reporting of information mitigates the noisy individual-level reporting of information [5].

Characteristics, such as how informed individuals are and their impact on colony survival, determines which individuals become leaders in colonies. The colony's survival is dependent on its queen's survival, who carries the new colony's genes [5]. Due to the queen's impact on the colony, she can be perceived as a leader. Scout honeybees can also be considered leaders, since they are responsible for initiating the departure of the daughter colony from the mother colony, make the choice of a suitable nesting cavity, trigger the colony's takeoff to the new nest site, and steer the colony during its flight [5]. The scout's knowledge about the colony state inside the hive, the weather outdoors, and the selected nest site, enables them to lead the colony. Patrol ants have similar outdoor knowledge, as they are the first to leave the nest in the morning. Patrollers search the nest mount and foraging area, as well as choose and inform (i.e., lead) the day's foraging directions to the respective foragers [48]. Other ants exhibit leadership roles, such as those who have found food and recruit others via tandem-running, where they lead others to the food site [49]. Some ants that are committed to a new nest site will assert dominance over passive adults by picking them up and carrying them to the site location [50]. Dominance is also exemplified in first virgin queen honeybees that pipe on the combs to transmit messages to the colony, which causes the workers to cease instantly all movement for the duration of her signal [5].

Colonies sometimes experience undesirable emergent behaviors. A honeybee colony can experience split decisions, which arise when it attempts to normally take off towards a chosen new nest site and fails to move, because half of the colony is supporting one choice, while the other is supporting another [5]. The honeybees will resettle and debate further in order to come to an agreement. Split decisions are wasteful and potentially fatal to honeybee colonies, since the members are exposed to risks associated with the outside environment throughout the decision-making process.

The roles suggest that a colony can influence individuals' actions in order to maintain behavior. The most common colony need that influences individuals' actions are nutritional. Receiver honeybees will only accept water from forager honeybees if they require it for themselves or to pass on to other honeybees [14]. Thus, the forager's response to finding and unloading water is regulated by the colony's need. Honeybee foragers perform two dances to reflect the needs of the colony. The waggle dance results in the recruitment of more foragers, while the tremble dance results in the recruitment of more receivers [14]. The nutritional needs of an ant colony also influence an ant's decision to leave a pheromone trail to a food source [14]. A change in the rate of forager return to the nest translates to the colony's need to send more foragers out [48]. Worker honeybees will change their interactions towards their queen in preparation for departure from the mother colony. The workers will begin to show mild hostility towards the queen by shaking, pushing, and lightly biting her, all of which is intended to help her lose weight so she will be able to make the journey to a new daughter colony [5].

The colony's feedback loop behavior is used to gain or mitigate support. Scout honeybees that produce dances for a particular site repeat their dances in order to convert neutral individuals into supporting individuals for the same site [5]. Scout honeybees that do not promote enough positive feedback for an optimal nest site may fail to recruit others, which may result in a suboptimal decision [5]. The selection of a lesser quality site may also occur if the scout honeybees provide positive feedback too late into a debate. A lesser quality site may have gained supported simply because it was entered into the debate earlier and had more time to gain support [14]. A naturally occurring negative feedback loop attempts to mitigate support for poorer sites. Scouts that were supporting individuals become apathetic voters after a duration of time and rejoin the pool of neutral scouts [5]. Determining how to balance positive and negative feedback

loops is necessary in order to embed this method in future human-collective systems.

2.0.2.2 Definition of Robotic Colonies

Robotic colonies are decentralized systems [51] composed of numerous entities (>50) who exhibit unique roles or states, such as foraging, which adapt over time to maintain consistent states in changing conditions [52]. Colonies use a centralized space for sharing information, such as inside of a nest, or embed information into the environment, such as ants depositing pheromones to communicate a route from a food site to the nest [53]. Colonies use strategies, such as positive feedback loops, to gain a majority consensus in order to change its behavior [54]. Recruitment for a change will continue to increase until a quorum has been reached. The colony will transition into a decision-making state once the quorum is reached. Different strategies, such as the honeybee waggle dance [5], will be implemented to reach a consensus.

2.0.2.3 Visualization Transparency on Human-Colony Systems

Less existing research investigated visualization transparency aspects in relation to human-colony systems. Abstract visualizations were designed to convey similar system features provided in traditional visualizations. The radial visualizations promoted perception (SAT Level 1) by displaying mission [55] and colony state information [56], such as the direction the individual entities left the hub to explore targets. Predictions of future colony headings (SAT Level 3), provided by elongating the radial display in the direction of more colony support, aided operator actions. Visualizations providing predictive information are needed to achieve transparency for human-colony systems,

regardless if the visualization is traditional or abstract. Additional human-colony based system evaluations are needed to establish a broader understanding of the influence of visualizations on human-colony system behaviors and performance.

2.0.2.4 Influence of Transparency on Human-Colony Interactions

Only one colony based evaluation assessed operator influence level and information reliability during a best-of-*n* decision-making task [56]. Operators placed beacons in the environment to attract support at particular locations. The direction of the individual entities was communicated to operators using a radial display surrounding the hub. Low operator influence resulted in high performance when reliable information was provided, while high influence was best when inaccurate or incomplete site information was provided. Further analysis is required to determine if less operator influence can be achieved when there is imperfect communication. More human-colony based system evaluations are needed to establish a broader understanding of the influence of different system design elements, such as control mechanisms, on human-colony interactions.

2.0.3 Collectives

A biological example of collective behaviors is provided in order to understand how collectives encompass both spatial swarm and colony behaviors. Only one evaluation assessed the influence of transparency on human-collective interactions and emphasized the need for more evaluations, such as the one described in Chapter 3.

2.0.3.1 Definition of Robotic Collectives

Collective robotic systems exhibit biologically inspired behaviors seen in spatial swarms [3], colonies [4], or a combination of both [5]. The honeybee colony searching for a new hive location example described in Chapter 1 exemplifies collective behavior. The daughter colony transitions from colony behaviors to spatial swarm behaviors when it initially flies to a nearby tree branch, where the daughter colony waits while scout honeybees search the surrounding area for a new hive location. The honeybees resume the colony behaviors during the consensus decision-making process in order to choose the best site. Once a decision has been reached, the colony transitions back to spatial swarm behaviors during the flight to the new nest site location.

2.0.3.2 Influence of Transparency on Human-Collective Interactions

Thus far, only one collective evaluation has investigated how humans can influence collectives based on two decision support models compared to direct control of the collective to achieve the same task. This same effort investigated the influence over the collective's decision-making performance with and without a human operator [57]. Four collectives, consisting of 200 individual entities each, needed to make multiple optimal sequential decisions. The sequential best-of-n decision-making model, that compensated for environmental bias, without an operator reached consensus slower, but made 57% more accurate hard decisions compared to the sequential best-of-n model that only assessed a target's value. The addition of an operator using the environmental bias compensated model required less operator influence and achieved 25% higher accuracy for the hard decisions. Further details are provided in Chapter 3, since this

dissertation extended on these results for transparency in such systems.

2.1 Transparency

Transparency represents the means of providing insightful information from the machine to the human operator and vice versa. Providing too much or too little transparency may overload or underload respectively, the human operator and negatively affect desired outcomes, such as performance. The use of various design strategies to integrate transparency can help improve the system's overall effectiveness.

2.1.1 Transparency Definitions

The most common robotics related transparency definition is "the quality of an interface to support a human operator's comprehension of an intelligent agent's intent, performance, future plans, and reasoning process" [39]. The three level Situation awareness-based Agent Transparency (SAT) framework was developed as a guide for achieving transparency. The SAT model leverages the human operator trust calibration 3Ps model (purpose, process, and performance) with performance history [58, 59], Endsley's situation awareness model [60, 61], and the Beliefs, Desires, and Intentions Agent framework [62]. Endsley's situation awareness (SA) model encompasses the perception, comprehension, and projection of future states of an environment and the actions to be taken.

At level one of the SAT framework agents communicate their current status, actions, and plans. The agent's 3Ps, desires, and intentions inform the human operator's perception of the system's current and future plans. The agent communicates its reasoning process and potential constraints or affordances to the human operator in order to sup-

port future action comprehension in SAT level two. SAT level three communicates the agent's future outcome projections, uncertainties, limitations, the likelihood of success, and performance history to the human operator. Teamwork transparency and bidirectional communication were added to the updated SAT framework in order to maintain a mutual understanding of each teammate's responsibilities and interactions [63].

Other transparency definitions, across domains, describe transparency as the communication of information regarding the machine's abilities [64] and capabilities [65]. Transparency has been described as a process [66], method [67, 68], mechanism [69, 70], property [71], or emergent characteristic [72] that provides information or explanations [73] to a human operator in order to develop accurate mental models of the system. The type of information provided, known as information transparency [74], includes what [75] the human operator or machine is doing [69] and why a particular task is being conducted [76]. Functional transparency [74], or seeing through a system [72, 77, 78], addresses how a task is accomplished [79].

This dissertation defines *transparency as the principle of providing information that is easy to use [77] in an exchange between a human operator and collectives to promote comprehension [69, 75, 80] of shared awareness [68], intent, roles, interactions [63], performance [39], future plans, and reasoning processes [81, 82].* The term "principle" is used to describe the process of identifying *what* factors affect and are influenced by transparency, *why* those factors are important, and *how* to design a system to achieve transparency.

2.1.2 Factors of Transparency

Designers of human-machine systems must consider the human operators, machines, and the interactions between them in order to understand why transparency is a critical

principle. Each entity possesses unique individual and shared factors. Identifying and understanding how these factors influence the perception, processing, and projection of actions during tasks is crucial. Factors associated with interactions between human operators and machines will differ depending on whether the interaction is *direct*, such as physically driving an automobile, *indirect*, for example via a computer interface, or manipulating the operational environment, such as obstacle removal. This dissertation investigates the direct and indirect interaction factors.

A human-machine system transparency factor concept map, presented in Figure 2.2, was created after reviewing the human-machine system transparency literature. The Google Scholar and Oregon State University Valley Library search engines, with the keywords "transparency and human machine" as well as "transparency and human robot", were used for the search. During the literature review, information related to transparency was transcribed and sorted by reoccurring themes (e.g., performance, usability, trust, and explainability). Factors that were identified can be used to measure the effectiveness of design choices intended to achieve transparency. The factors can either be *influenced by* transparency or *affect* transparency. The concept map identifies factors related directly (solid lines) and indirectly (dashed lines) to transparency, which were referenced by the supporting literature. The direct factors had immediate connections related to transparency, such as transparency has been described as observable, whereas indirect factors typically influenced other factors, for example credibility impacts trust, which impacts transparency. The three indirect factors, *feedback*, *state*, and *prompts*, were concatenated into one particular nomenclature in order to provide clarity.

Fourteen direct factors surround transparency, of which, the four highest total degree (number of in degree + number of out degree) direct factors are the focus of this dissertation (dark blue in the figure): performance, usability, trust, and explainability.

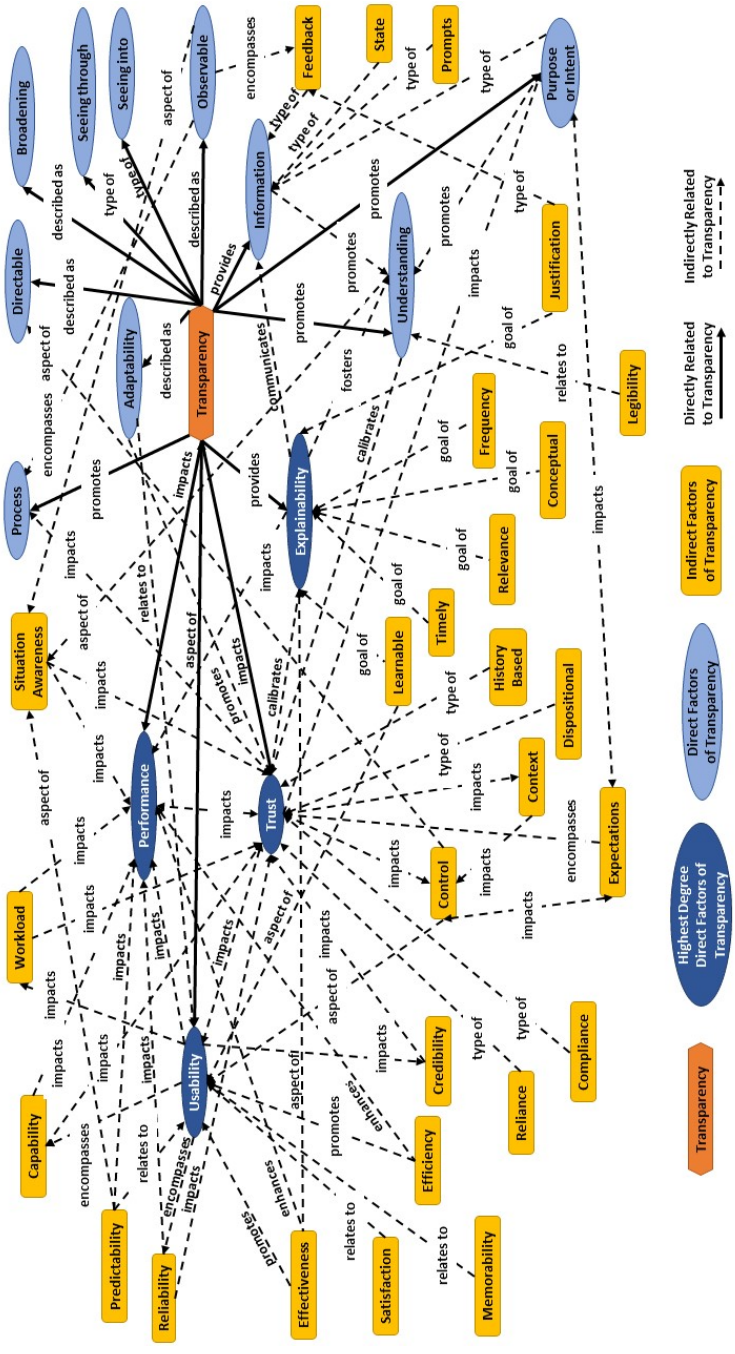


Figure 2.2: Concept Map of human-machine system direct and indirect transparency factors [2].

The information, purpose or intent, and understanding factors were not considered high degree direct factors, because explainability uses information, such as intent, via explanations, to communicate and promote understanding of each entity type. Many indirect information factors were considered methods of achieving transparency.

The factors and how they interact with one another are provided in order to address the "principle" transparency aspect. All the direct factors are defined, and the associated indirect factors are indicated. Related research outside of robotics is presented as "Other Domains", since the transparency research originates from other system domains. An overall summary concludes each chapter in order to highlight relevant direct factor elements and applicability for supporting human-machine system design.

2.1.2.1 Performance

Performance is the ability of a human operator or system to produce an output when executing a task under specific conditions [80]. Performance can be measured subjectively (i.e., surveys) or objectively (i.e., psychophysiological responses or task completion time). Teamwork can increase performance [83], emphasizing the need to make effective interactions between team members. Various direct and indirect factors, in Figure 2.2, impact performance, including explainability (i.e., the form of information exchange and communication), trust, SA, workload [72, 84], as well as system reliability, predictability, and capability [39].

Other Domains The effects of explainability, via system reasoning on human operators' trust, workload, and performance were evaluated for various aircraft and autonomous automobile applications [81]. Transparency positively influenced perfor-

mance and calibrated trust, but produced higher workload and longer decision-making times. This result can be attributed to the amount of information provided. System transparency impacts on performance were assessed for individuals and groups using collaborative information-visualization environments [75]. Users had higher net productivity when the design accounted for transparency intentionally. Cost-benefit analyses of various visual support systems, using the Human-Automation Collaboration Taxonomy framework, evaluated how information transparency impacted decision-making performance in a missile strike coordination application [74]. The framework was useful when developing cost functions related to money, time, performance, and safety due to the differing visual support system designs.

Robotic Domains Attributes of decision and control environments [84] impacted real-time trust and control allocation strategies for collaborative human-robot teams [85]. Decrements in reliability affected subjective performance assessments, trust, as well as the frequency and timing of switching between autonomy modes. Multiple experiments using the Autonomous Squad Member and Intelligent Multi-Unmanned Vehicles Planner with Adaptive Collaborative/Control Technologies examined the effects of transparency level on operator performance [86], calibration of trust [64], workload [87], and perceived usability [88]. Generally, performance benefited and trust was appropriately calibrated with better transparency; however, the presentation of uncertainty information, in SAT level three, did not improve performance consistently.

Summary Performance is an effective metric to assess transparency effects in human-machine systems. There is a high association between performance and other transparency factors (eleven total degrees in Figure 2.2): three of which are the high degree

direct factors, seven are indirect factors, and the remaining is transparency. The high number of interactions imply that performance is necessary for understanding and implementing transparency effectively. Performance is an outcome of interactions, including direct and indirect factors, as well as the task and environment. Real life scenarios make determining exactly what causes performance changes challenging. Studies that attempt to duplicate realistic use cases can help designers understand the combinations of factors that result in certain outcomes and inform identifying desired performance ranges that are robust to variable contexts and environments.

2.1.2.2 Usability

Usability is a multifaceted quality that enables human operators to achieve desired goals in a manner that can be anticipated, easily learned, and does not cause confusion or hinder progress [89]. Various direct and indirect factors, from Figure 2.2, are associated with usability: learnability, efficiency, effectiveness, memorability, satisfaction, control, predictability, and transparency [78, 89, 90]. A system designed to leverage prior human operator knowledge, from mental models or schemas, can be easier to learn, which may make using the system more efficient, effective, easier to recall, and pleasant [90]. Control is the authorization an operator or system has over a particular task [78], where the human operator and system may assume various roles, such as a supervisor or teammate, depending on the situation's context.

Transparency has been described from a usability perspective as observable, directable, adaptable, and broadening [39]. A system is observable if feedback is provided to the human operator and machine about a process [39, 91]. The human operator or system, dependent on the feedback received, may decide to re-direct or mod-

ify resources, activities, priorities, and assumptions in order to explore other solutions [39, 91, 92]. Human-machine system adaptability is the system's ability to adjust to situations and provide support to a human operator in order to accomplish tasks, which is necessary when operating in variable environments. Lower levels of automation may require the human operator to have more influence over tasks compared to higher levels of automation, where the responsibility lies more with the machine [93].

Other Domains Transparency's impact on government website usability for characteristics, such as website layout, intuitive menu systems, site maps, and search tools has been analyzed [94, 95]. The human operators' overall web knowledge impacted usability and transparency [94]; however, trust was an important factor when human operators were critical of the provided information [95]. A design intended to improve usability of security software actually resulted in issues due to a lack of transparency, as human operators did not know when or how to make security related decisions [96]. Transparency increased usability in web authentication systems; however, interactions between the users and system decreased, which caused the users to become confused and develop a lack of trust in the system [97].

Robotic Domains Transparency has, in telerobotic systems and human-robot teams, reduced operator workload, facilitated operator comprehension, mitigated errors [98], improved usability [99], and positively impacted perceived system dependability [79]. Identifying and affirming which interface symbology supports developing appropriate mental models can improve the overall usability of human-robot interaction design [100]. Robots that increased their transparency levels improved human operator performance, but the perceived usability did not improve [88]. A system that provides

increased transparency can be perceived as more complex, which can negatively influence perceived usability and workload (both subjective and objective) [64].

Summary Usability influences the human operator's perception of a system. Figure 2.2 indicates that usability has a higher number of interactions with the transparency factors (fifteen total degrees) compared to performance, making usability a necessary factor for understanding and implementing transparency. Two of the fifteen total degrees are the high degree direct factors, twelve are indirect transparency factors, and the last is transparency. The direct and indirect factors related to the system's characteristics, such as control, influence the operator's opinion regarding the system's usability. Making hardware or software more reliable, can improve usability; however, because perceived usability is subjective, the operator's perception and satisfaction may contribute to overall system usability. Designers must consider objective and subjective aspects of usability to understand how factors affect operators and identify what transparency modifications can be implemented to maximize the system usability.

2.1.2.3 Trust

Trust is a psychological state [101] and sentiment [102] that compares an operator's willingness [68, 103] to be vulnerable [80] and confident [104] regarding the expectations [105] of honesty, fairness, care, and responsibility [83]. Trust emerges when the human-machine system can commit, reflect, and adapt to recommendations, actions, and decisions [83]. Various types of trust exist, including *compliance*, *reliance*, *dispositional*, and *history-based*. *Compliance* arises when an action occurs after receiving a cue, while *reliance* is avoiding an action because there is no cue [106]. The operator's attitude

towards a system provides essential information when designing human-machine systems [105]. Prior interactions (*history-based* trust) or no prior interactions (*dispositional* trust) contribute to the operator's attitude towards a system [71].

Trust beliefs, intentions, and actions must be considered when designing systems. An operator will take control of the system if the system's action is perceived as inadequate [77, 80, 107]. Trust intention and reliability [108, 109] help identify appropriate expectations and potential sources of uncertainty [110]. Trust in the system may increase if expected responses emerge from particular commands; however, when the system responds unexpectedly, trust may diminish. Understanding how much each entity, the human operator and system, can rely on one another during a task is essential for trust calibration [80]. Human operators who have too much trust, or overly rely on a system typically misuse it, while not trusting the system can lead to disuse [111].

Three factors, in Figure 2.2, were identified for developing trust: *purpose*, *process*, and *performance* [58]. The *purpose* corresponds to the system's intended use, the *process* represents the human operator's understanding of the system's logic, while the system's observed behavior represents *performance*. Trust is influenced by the system's credibility and capability [80], individual human operator differences [58], such as workload and situation awareness [112], and teamwork aspects, such as respective roles, division of labor, shared environmental awareness, and context [63, 110]. Transparency is not expected to improve trust, but can calibrate trust appropriately, depending on the context [39]. For example, increased transparency that reveals information related to system reliability, can cause the human operator to trust the system less.

Other Domains Participation, transparency, and communication between employees at two different companies were examined in order to understand the impacts on coop-

eration and trust [66]. Different communication mechanisms, such as blogs and wikis, improved the intra-organizational information transparency. Trust in a coffee production machine [77] and in a private webmail system [113] was rated lower for higher transparency levels due to increased perceived complexity. Two transparency factors, readability and organization, were altered from various code scripts [80]. Readability of the code led to the highest trustworthiness, which alludes to the importance of designing interfaces that are easily understood. The relationship between a human operator's expectations and system output, in an online peer assessment system, provided evidence for a bell-shaped relationship between transparency and trust [105]. The bell-shaped relationship insinuates that an opaque or highly transparent system may produce equally low levels of trust. Interfaces, used for emergency landing in commercial aviation produced higher trust and utility with improved transparency [114].

Robotic Domains Transparency effects on human operator's trust [115], situation awareness [67, 87], and workload [71] were assessed for the Autonomous Squad Member. Higher transparency levels improved situation awareness, but did not always improve trust or workload. Presenting the system's uncertainty information may have introduced ambiguity, lack of relevance, and incomplete knowledge of the system's operational capabilities, causing mixed human operator responses [81]. The Intelligent Multi-Unmanned Vehicles Planner with Adaptive Collaborative/Control Technologies was used to evaluate the impact of transparency on performance, trust [116], workload [64], situation awareness, and reliance [88]. Performance improved with increased transparency; however, trust, did not improve with increased transparency. Trust was examined by determining whether automation reliability and automation transparency had different implications on human reliance behaviour and mission performance [117].

Higher reliance rates and short response times were promoted using either reliability or transparency based trust. Human-robot trust was investigated in a collaborative table clearing task [118] as well as a tactical task [84], in which the robots learned the desired actions with the intention of increasing the human operator's trust in the system. Periods of low reliability early during the interaction phase that included real-time feedback had a detrimental impact on trust [85]. During an unplanned robot encounter reactive planning techniques were used to build transparency [82], which improved overall understanding of the system. Different transparency levels of conflict detection and path re-planning were provided to users tasked to identify and attack hostile targets and reroute unmanned aerial vehicles' paths to avoid conflicts [119]. Higher transparency increased operator dependence on the automation and increased trust.

Summary Trust is highly complex, and various factors can affect trust or are influenced by it. Trust has the highest association with the transparency factors (twenty total degrees) compared to performance, usability, and explainability, as shown in Figure 2.2: three high degree direct factors, sixteen indirect factors, and transparency. Trust must be considered in order to understand and implement transparency effectively. All three stages of human processing: perception, comprehension, and projection are influenced by trust. Some factors that impact trust are related to system capabilities and how it is designed, while other factors are characteristic of the human operator. Research demonstrates that transparency can improve or reduce trust. Designers must consider strategies, such as providing supplementary information, to improve human operators' confidence in the human-machine team, and consider how the system design influences trust. Building the human operator's trust and the system's ability to execute tasks correctly will be critical for future use.

2.1.2.4 Explainability

The clarification or justification of actions is the explainability factor, which was touted to promote trust, positive and constructive interactions, as well as transparency [120]. Principal components of providing effective and comprehensible explanations are the human-machine systems facts, plans, and goals [121]. Figure 2.2 identifies five goals from an explanation framework for dynamic problems with high levels of uncertainty: transparency, how answers are reached; justification, why an answer is correct; relevance, why a question is relevant; conceptualization, clarify ideas or notions; and learning [122]. Explanations that incorporated components, such as justification, improved the automated collaborative filtering systems' acceptance by increasing human operator involvement in the reasoning process [123]. Classifiers centered on what, why, and when actions were going to be performed by a robot were used to provide interpretable explanations of the robot's needs, behaviors, and intentions [121].

Other Domains Understanding how explanations impact human operators' perception of control over the system revealed that a lack of explanations resulted in humans exerting more control [124]. The possibilities, challenges, and effects of explaining system reasoning on the human's trust, workload, and performance were examined for fighter aircraft, air defense, and autonomous driving [81]. Explaining the system's behavior and inferences improved collaboration with the human. Multiple explanation types were evaluated to determine which were effective for human operator understanding and performance [125]. Explanations indicating why the system behaved in a particular manner produced higher understanding and stronger feelings of trust.

Robotic Domains An experiment in which a simulated autonomous agent controlled a simulated unmanned ground vehicle examined how the simulated robots learned behaviors using a case-based reasoning approach [126]. The robot provided simple, concise, and understandable explanations of why a behavior was executed, resulting in higher performance. The effects of transparency on the attribution of credit and blame were assessed using a delivery robot [73]. Transparency was dependent on the similarity between the robot's explanation of its actions and the participants' knowledge. Accountability of a swarm's actions (i.e., what the swarm did at particular times) was needed by fire and rescue personnel conducting investigations [127]. Challenges for human-companion robot collaboration, during plan-based and action-related problems, were investigated [128]. Plan-based strategies identified the importance of making robot behavior legible, which is related to the robot's understanding of their respective responsibilities and contributions in a joint plan. Human operator trust was reestablished when working with a turtlebot, which made and apologized for mistakes, only when the robot's explanation or apology occurred in a timely manner [129]. Novel algorithmic explanations, in two human-robot studies, improved decision-making and team performance, as well as improved transparency and trust [130, 131].

Summary Explainability is an assessment of prior, current, and future information that must be understood by operators and system entities within human-machine systems. Explanations are the mechanism to provide transparency. Forms of explanations include words or figurative icons. Explainability is a bidirectional transparency factor. Figure 2.2 shows a high number of interactions between the explainability and transparency factors (eleven total degrees): two are the high degree direct factors, eight are indirect, and the last is transparency. Explainability and the other factors it encom-

passes (e.g., information, understanding, purpose or intent) are necessary for effective understanding and implementation of transparency. Providing an appropriate amount of information is important, as well as providing relevant, useful, and timely information. The working environment may impose restrictions on explainability, such as limited bandwidth, that may affect the human-machine team negatively. Explainability is subjective, but has objective indirect factors of transparency.

2.2 Design to Achieve Transparency

Three well received and highly used design methods can be implemented to optimize human-machine system transparency and desired outcomes. The designer can 1) *provide* system features, such as providing feedback, 2) design systems using specific *guidelines*, for example, Gestalt principles [37], or 3) *train* the human operators and system, which is especially important for complex systems. An ideal system uses a combination of all three methods (provide, design, and train) to ensure optimality.

2.2.1 Providing Characteristics

Four criteria to *provide* in a design for transparency were identified: status (S), feedback (FB), planning mechanisms (PM), and engagement prompts (EP). Three of the four criteria (status, feedback, and engagement prompts) are directly related to the information factor, identified in Figure 2.2. The planning mechanism does not appear in Figure 2.2; however, it encompasses various direct and indirect factors.

How the indirect transparency factors can be provided (P) or assessed (A) for each criterion and the relationship with respect to the direct factors, addressing the "princi-

Table 2.1: Transparency factor information used when Assessing (A) or Providing (P) a system Status (S), Feedback (FB), Planning Mechanisms (PM), and Engagement Prompts (EP) [2].

		Direct Factors															
		Performance				Usability				Trust				Explainability			
		S	FB	PM	EP	S	FB	PM	EP	S	FB	PM	EP	S	FB	PM	EP
Indirect Factors	Control			A	A	P	P	A	A	P	P						
	Reliability	A,P			A	P	P		A	P			A				
	Workload	P	P			P	P			P	P						
	SA	P	P		P						P					P	
	Capability	A		A,P	A,P	P											
	Information													P	P	P	A,P
	Understanding													P	A	P	P
	Context										P	P	A,P				
	Process										P	P	P				
	Timely					P								P	P		A,P
	Effectiveness		P			P											
	Efficiency				P	P											
	Expectations											P	P				
	Learnable						P								P		
	Predictability		P				P										
	Purpose										A,P	A,P					
	Reliance										P		A,P				
	Compliance											A					
	Frequency													P			
	Justification														P		

ple" aspect of the transparency definition, are provided in Table 2.1. The associations in Table 2.1 were developed by one researcher and supported by the literature review. Particular keywords related to the criteria (e.g., status/state, feedback, plan, prompt, alert, and alarm) were used to identify whether an association (e.g., provide or assess) existed between the direct and indirect factors. Initially, information about the associations was transcribed and sorted by criteria type. The associations were classified as *provide*, which used the words provide or a similar synonym, or used the word *assess*, which used the words assess or a similar synonym. For example, if an evaluation focused on understanding how providing transparency impacted workload via a status, the researcher identified what direct factors were associated with workload from Figure 2.2; in this case, workload impacts performance and trust. A "P" was assigned under the status sub-columns corresponding to the direct performance and trust factor columns for the workload row, which are identified in Table 2.1. A designer can use the information from Table 2.1 to determine which indirect factors may be used to evaluate design decisions. System reliability information is an example of a metric that can be used to evaluate a status to promote transparency. Each criterion is defined, the associated advantages for human-machine systems are identified, and how the criteria can be leveraged to assess transparency at the three SAT levels is provided.

2.2.1.1 Status

Status incorporates the *what* aspect of the transparency definition by providing the state of the human operator or system at a particular point in time. The current state's status is necessary in order to evaluate performance of the human operator, machine, and human-machine system with respect to the human operator's and system's available

capabilities [80]. Various types of indirect factor information, see Table 2.1, can be provided in a status to help determine whether changes to the strategy by the human operator or system must be initiated to accomplish a task.

SAT Level 1 Operators working with teammates, humans, or system entities, tend to lose awareness of environmental and system changes [132]. Loss of awareness is problematic and can result in undesired consequences; hence, providing state and action [133] information can improve teamwork efficiency and reliability [98], as well as indicate what information is missing, incomplete, or invalid [112]. Providing a status of the human operator's, system entities', and system's states can indicate where errors are occurring, aiding in the development of control strategies to mitigate ongoing errors, which can maintain an appropriate level of trust between the entities. Presenting status changes can improve the human operator's effectiveness, by alleviating the time and effort devoted to integrating information and drawing conclusions about a situation, which may be impeded by interruptions [134].

Information, such as tracking the task's completion progress, is helpful for understanding progress in relation to goal achievement [133]. The timing of status messages is crucial and requires consideration of the operator's capabilities, system limitations, the task, and the environment in which the human-machine system operates [112]. Providing too many messages can overload an operator and the frequency of messages can result in overtrust or mistrust, all of which can lead to substandard performance.

SAT Level 2 SAT level 2 provides information related to the system's reasoning process or motivation, which are types of feedback and are discussed in Chapter 2.2.1.2. Understanding a machine's functional status can calibrate and maintain an appropriate

level of human operator trust in the machine's reliability and ability to support a task [72]. Providing status changes can provide a deeper explanation of what those changes were, as well as why and when they occurred [135].

SAT Level 3 Providing state projections can facilitate the human operator's understanding regarding the current action consequences on the system's future state [133]. The human-machine system can leverage the projections to develop new, or revise existing, actions and strategies in order to ensure task completion. Predictive displays that show state characteristics, such as a robot's projected position, can alleviate latency effects, which limit a human operator's ability to influence a system's actions [136].

2.2.1.2 Feedback

A feedback mechanism, identified in Table 2.1, can provide explainability, via descriptions, that justify or provide insights into actions, uncertainties, reliability of recommendations, and supplementary information from the human operator or machine system [81]. The *why* aspect of the transparency definition is addressed by providing explanations, which is a form of feedback. Multimodal feedback incorporates various sensory channels (e.g., visual, auditory, and tactile) in order to optimize communication with a human operator and has been deemed useful in human-machine teams [36].

SAT Level 1 Accessibility to raw data, known as seeing through the system, can enable the human operator to feel in control [72] of a situation and assure system information accuracy [112]. However, the human operator's workload may continue to increase with additional information, exceeding their capabilities and potentially limiting

the system's perceived usability and trust in it [64]. Lack of feedback, especially in cases where the system has more control over a large number of tasks, compared to a human operator, can decrease a human's situation awareness [132]. Information presented in formats that leverage human cognitive processes can positively influence trust [112].

Various studies assessing transparency, such as the system's purpose and range of applications [58], have resulted in improved trust due to better trust development [112], decision-making, and team performance [130]. Operators had higher task performance when effective explanations identified why the system behaved in a particular way [125]. Explanations of why system errors may occur increases reliance [137] and improves control allocation strategies between the human and machine [85], as well as mitigates blaming the machine [73]. Suboptimal sensory quality due to environmental conditions or the age of the information [138] can affect system reliability negatively. Providing timely feedback regarding the machine's awareness of environmental conditions, constraints, and task-related limitations may help the human determine what contributions to make in order to maintain task progress [110, 133].

The quantity of feedback provided to a human operator must accommodate human capability limitations and available decision time [139]. Providing explicit and implicit feedback, such as completion of tasks or interruptions, can support the human's understanding of the system's state [126]. Solely providing information to operators does not guarantee accurate perception and comprehension. Offering corrective and developmental feedback that highlights mistakes and provides suggestions for future actions can improve system performance [140]. Understanding how the system responds in different environments is necessary to provide corrective feedback to the operator [141].

SAT Level 2 "Seeing into a system" explains how the system collects and uses information. Revealing the system's rules and algorithms can improve human operator trust, if the human understands what is happening [58, 112]. Providing simplified system process feedback [58] and displaying intermediate results from the algorithmic process [112] can promote better human operator comprehension. Contextual information, such as system settings or environmental conditions [112] that influence the algorithm's response, enabled better trust calibration, improved performance related to time management [137], and enhanced understanding of system vulnerabilities [138].

Relevant feedback can help the human operator maintain control of the system, improve situation awareness, and alleviate workload [81]. Perceptions of trustworthiness can define reliability and lead to trust, perceived utility, and reliance; however, the perceptions depend on an understanding of human-machine expectations and how those interactions impact one another [110, 133]. Human operators use cost-benefit analyses to determine their trust in a system [58]. When expected benefits are violated, human operators instinctually attempt to find possible justifications for inconsistencies between the expected and actual outcomes [105, 142]. Explanations regarding the violation must contain causal information about how prior actions led to the current state, in order to facilitate understanding, acceptance, and trust [120, 123]. The designer must avoid introducing bias when suggesting evidence and provide an estimation of the suggestion's reliability [81] without causing cognitive tunneling.

SAT Level 3 Feedback regarding the system's strengths and weaknesses can educate human operators [123], which increases acceptance, understanding of information reliability, accuracy, and quality, as well as calibrates trust appropriately [81]. Tracking decisions during task execution, and providing performance information, can facilitate

the human-machine team's understanding of what modifications to make in order to improve task execution [112]. The information can update the human's mental models [133]. Feedback must be provided to both the human operator and machine, since trust is expected to fluctuate as the history of interactions and system performance changes [67]. Broadening feedback considers various "what if" scenarios based on numerous combinations of human-machine interactions [92] and may help guide the human operator and machine towards an optimal future desired outcome.

2.2.1.3 Planning Mechanism

Planning mechanisms encompass the allocation of resources and task assignments among an organization's members, such as the expectations of a human-machine team to fulfill a goal, identified in Table 2.1. The *how* transparency definition aspect is addressed by the planning mechanism, due to the various strategies associated with plan development and maintenance. Planning occurs at all mission stages and is necessary for human-machine coordination in order to maximize desired outcomes [128].

SAT Level 1 The anticipatory planning stage begins by establishing the overarching goal and purpose of the shared human-machine team [110]. Shared representations of the team's purpose and how each team member contributes to mission success are necessary to set realistic expectations [92]. During the decision-making process, roles are defined and acceptable behaviors and interaction expectations outlined, which alleviate miscommunication or misunderstanding [110, 112]. Team member control coordination can be identified only if the skills, potential strategies, and procedures are understood for a desired task [84]. Shared mental models enable teams to coordinate actions and

adapt their behavior to emerging demands via explicit or implicit coordination. A decrease in communication and coordination overhead result from shared mental models [133], which is advantageous in limited communication situations.

The human-machine system performance will be dependent on the compliance of each team member to adhere to the proposed plans [58]. Underlying assumptions in human-robot collaboration scenarios assume that 1) all actions executed by a human are relevant to a mission goal and that the robot understands the intention of the action, and 2) that the human will always accept assistance from the robot [143]. Expectations impact how human operators calibrate and maintain trust, which are frequently reevaluated due to changing circumstances [58]. The analytic process evaluates information using the system's prior knowledge and experience, while the analogical process develops trust based on rules and procedures. However, the core influence of trust is based on the affective aspect, which claims that people make judgments on the impression of what they feel [58]. The subjective perspective; therefore, hinders the validity of the traditional assumptions, challenging human-machine system designers to consider subjective differences and preferences during planning.

SAT Level 2 During mission execution the human-machine system undergoes a process that transforms gathered information into an understanding of the implications of the information received, and uses the information to accomplish a task [144]. The team members need to maintain shared awareness and perceive, comprehend, and act upon the information to ensure successful performance [145]. Deviations from shared awareness require the team to re-evaluate whether the initial plan must be modified to ensure mission success. The coordination required to ensure smooth transitions during in-processing planning will be impaired if there is a misunderstanding of the team mem-

bers' roles, functional capabilities, and limitations [146]. Providing a mechanism that enables the human-machine team to modify pre-existing plans and maintain shared awareness in order to adapt to emergent behaviors is necessary for mission success.

SAT Level 3 Ensuring successful execution of plans requires the human operator to understand the system's behavior, operational boundaries, and limitations [83]. Reactive planning can be flexible to unexpected behaviors that may arise during a task [128] and can support successful plan execution. Information regarding the team's interactions, such as deviations from previous plans and adapted responses [147], can aid supervisors in the human organization's leadership when re-tasking team members' roles and responsibilities. Effective communication strategies will need to consider plan explicability, how human operators interpret plans, and offer proposed courses of action [120]. The likelihood of success or failure, assessed from the human-machine performance perspective, must demonstrate how alternative plans affect mission outcomes. Comparing alternatives, using metrics, such as cost-benefit analysis, ensure selection of the best future course of action [120, 146].

2.2.1.4 Engagement Prompts

Out-of-the-loop issues, in Table 2.1, such as a lack of context, arise when a human operator becomes isolated from contributing to the human-machine system, and can be mitigated by engagement prompts. Engagement prompts are cues, alerts, or warnings that encourage the human operator's involvement in an attempt to bring the operator back into-the-loop. Engagement prompts encompass the three aspects of transparency (*what*, *why*, and *how*) by indicating to the human operator why they became disengaged,

what action must be taken to continue or return to the task completion, and identify different strategies that can be implemented to fulfill the task.

SAT Level 1 The updated SAT model incorporates bidirectional transparency to describe the transparency needed to support human-machine collaboration [63], which addresses the knowledge structures to facilitate the interactions between the human operator and system. No ideal interaction exists, but there is a range of possibilities that emerge from varying contexts and resources, including the system's opacity (inverse of transparency) [148]. Suitable information regarding the state of operation can change with transparency level [141]. Different questions regarding the environment will emerge as the state's representation changes. If a human operator becomes disengaged, even temporarily, reintegration into the loop will be challenging. Implementing system engagement prompts that remind the human operator to engage with the machine can improve the effectiveness of the interactions and calibrate trust appropriately. The timing of the engagement prompts is also critical, since the human-machine system's state can vary drastically, dependent on when a prompt is issued [34].

SAT Level 2 Proactive monitoring by the system can alleviate near misses when a human operator is disengaged or interrupted by other tasks or operators. The system can use feedback cues, such as lack of commands issued or psychophysiological responses, to determine if the human is aware of the system's state and what action to perform [149]. Depending on the human operator's state, the engagement prompt may be altered to assure an appropriate reaction. Alarms increase the likelihood of non-reliance [109], which affects perceived reliability. If the machine senses that a human is fatigued, issuing an alert may provide the appropriate salience needed for engagement; however,

if the human is frustrated, a prompt offering additional information may be useful [69].

SAT Level 3 A system can monitor task execution and alert human operators when failures have been detected or when a failure is likely to occur [133]. Different alerts can be used to notify the human operator of the severity of the failures. Significant system reliability changes, that may impact goal completion dramatically [135], such as low robot battery levels or malfunctions, as well as temporary actions needed to ensure goal success, may be communicated to the human operator by providing future projections. Designers must consider the repercussions of using different types of engagement prompts in order to avoid complacent behavior or alarm fatigue, which occurs when too many alerts are provided to the human operator causing them to ignore the alerts altogether. Past performance history can indicate which prompt types were more effective under specific contexts and identify optimal prompt engagement timing.

2.2.2 Using Design Principles

Design principles provide human-machine system designers a set of guidelines, derived from knowledge and experience of various systems, that can be used during the design process. Typical design processes begin at the problem development stage where operator needs and the system requirements are identified. Conceptual, preliminary, and detailed designs, also referred to as prototypes, are created to test whether the requirements and needs have been met. Designers will need to create multiple prototypes and evaluate them based on the needs and requirements. Integrating human operators during the prototyping phase is advantageous to capture interaction affects. Once a finalized design is achieved that meets the desired outcomes, the production

phase begins. This dissertation considered principles from the human factors and cognitive engineering domains that relate to displays and interfaces, such as perceptual, mental model, attention, and memory principles [150]. Each principle is defined and how the principle leverages transparency factors are provided at the three SAT levels.

2.2.2.1 Perceptual Principles

A human operator receives information via their sensory system, processes the information using cognition, and performs an action through their motor system, which produces a response. Perception is the ability to acquire information, from an entity or the environment, via different sensory modalities, such as visual, auditory, tactile, smell, and taste [151]. The process of integrating the individual stimuli together to formulate meaning is referred to as perceptual organization [1]. Perception is determined using bottom-up processes driven by the nature of the stimulation, as well as top-down processes, such as context and expectations. Human operators interacting with visual displays typically use vision to obtain information from a display, listen for information coming from the system, and influencing the system via a mechanism, such as a computer mouse; therefore, the focus of this dissertation was on these modalities.

SAT Level 1 System displays must be visible, legible, or audible [150] if an operator is expected to interact and understand the system's current status, actions, and plans. Visibility can be achieved by designing the system to be detectable under all viewing conditions [1]. The brightness, perceived intensity of a light source, and legibility of the display, such as the contrast of text or icons on the visualization background, as well as the illumination from the working environment, which may be variable, must

be designed to ensure information is visually salient to the human operator. Visual perception is impacted by various factors including, color, depth, and motion perception, as well as pattern recognition. Using color as a strategy to convey intent of the system to the human operator, such as red to denote danger, is an effective design tool unless the human operator is color blind. Displaying information redundantly via other strategies can help mitigate situations where human operators are incapable of perceiving all stimuli. Designers must avoid absolute judgements, which is a human's capability to judge the value of a variable [150]. Humans can only discern up to seven levels of coded variables accurately. Gestalt principles of object perception can aid designers to create stable, consistent, and simple interpretation of visual interfaces, such as orienting similar objects or information near one another, or using larger font sizes [152].

SAT Level 2 A designer of human-machine systems can support human operator comprehension of the system's reasoning process by using a top-down process. Humans perceive and interpret information in accordance with their expectations [150], which are formulated from previous experiences, biases, and heuristics [153]. For example, aligning information to read from left to right and top to bottom [154] may be an effective design strategy that uses top-down processes, if the human operator is from the United States, since that is the common text formatting in that country. Presenting information redundantly through alternative forms, such as voice and print or color and shape [150] increases the probability that the human operator comprehends the system's reasoning process and corresponding future actions.

SAT Level 3 Studies that have provided uncertainty information to human operators in various types of human-machine systems have found conflicting results. The ad-

dition of uncertainty has had minimal positive influence on performance or SA [86], and sometimes caused trust to decrease [87]. Misinterpretation of the uncertainty information may be a reason why this behavior has been observed. Further evaluations are needed in order to determine how presenting uncertainty information to participants impacts desired metrics. Describing information in a familiar way to the human operator and similar to real world conventions is one design strategy that can be implemented to help mitigate misunderstanding and improve the human operator's comprehension of current actions on future system states.

2.2.2.2 Mental Model Principles

Mental models are structures that reflect a human operator's or system's understanding of artifacts or concepts, created from their experiences interacting with those objects or notions [155]. Mental models are continuously used to help human operators or systems during learning, problem solving, and rationalizing behavior [156]. Mental models are dynamic and develop over time or can be carefully formed through training [157]. A human-machine system can develop shared mental models where human operators and the system describe the roles and responsibilities of the teammates, explain particular behavior and coordinate their actions, and predict adaptive actions that transpire from emergent behaviors [110, 158]. Accurate shared mental models can improve human-machine interactions and result in better performance.

SAT Level 1 Elements displayed on a visualization, via an icon or picture, to a human operator must be representative of that particular element and behave in a similar manner [150] in order to ensure consistency between the human and system. Design prac-

tices that implement measures to improve consistency can help convey the system's intent and mitigate errors caused by human operators due to failures in judgement (mistakes) or failures to execute necessary actions (slips) [159]. Dials, levers, or buttons within a visualization must consider the principle of moving parts, in which dynamic movement must be consistent with the expectations of human operators' mental models [160]. A dial rotated to the right and a lever pushed up are expected to increment the associated value proportionally.

SAT Level 2 Providing information to the human operator regarding how a system processes information and acts upon that information will restructure their mental model, which in turn will provide better comprehension of the system. Designers that do not provide sufficient information to human operators hinders their ability to evaluate whether the system is performing appropriately [161] and potentially reduces the number of interactions between the human and system.

SAT Level 3 Mental models are often incomplete, can be easily confused, and are structured based on inaccurate information or inappropriate analogies [162]. Poorly structured mental models can create a lack of confidence in the human operator's ability to identify or describe a problem that is not well understood [163]. The designer can help restructure appropriate mental models by providing feedback about the system's limitations and uncertainties. Useful feedback will aid the human operator's understanding of what current actions may do to influence future system state projections.

2.2.2.3 Principles Based on Attention

Attention represents the human operator's ability to concentrate on and process information from their sensory modalities, at a particular point in time, for a range of elements within an environment [163]. Various types of attention exist including: selective, which directs concentration to particular areas of an environment; focused, narrows the field of concentration to a small size in order to avoid distractions; divided, widens the field of concentration to accommodate multiple areas of interest; and vigilance, which sustains attention on a particular area of interest for a long time duration [157]. A human operator's attentional ability is influenced by their level of arousal during a task. The Yerkes-Dodson law suggests that too little or too much arousal will produce poor performance [164]. High arousal may cause perceptual narrowing, which leads to narrowing of attention, a keyhole effect example [165]. Human operators experiencing the keyhole effect are described as looking through a soda straw at the environment, which causes difficulties with perception, comprehension, and projection.

SAT Level 1 The principle of conspicuity refers to how well an element attracts attention [1]. Designers can implement various strategies to increase conspicuity, such as placing an element in a location that receives a high focus of attention, or increasing the element's salience to draw attention to a particular location. Designers must be mindful of how much salience is introduced to the human operator and how frequently. Increased exposure of saliency objects may reduce the significance of the associated message being conveyed. Interface, or interaction components, that must be mentally integrated can be placed closer to one another [150] or grouped together [154] in order to expedite processing and mitigate an undesired division of attention.

SAT Level 2 The quality of information provided regarding the system's reasoning process is more beneficial to human operators than the quantity of information. Quality is discerned not only by what the information is, but also by how easy it is to access. Designers can minimize information access cost, which is the time it takes the human operator to find the information they want, by placing more important things at the center of the human operator's field of view [150] or emphasizing important information by increasing saliency, such as the use of larger font size [1]. The best modality to convey information to the human operator is dependent on various aspects including, the information complexity, time to deliver the message, whether immediate action is required or not, and the environmental work conditions [1]. Auditory messages are simple, short, require immediate action, and are useful in working environments that have poor visibility as well as tasks that require the human operator to move to different locations within the environment. Conversely, messages that are long and complex, do not require immediate action, and can be implemented in noisy environments where the human operator remains in one position, benefit from a visual modality.

SAT Level 3 The intelligibility principle seeks to provide clarity regarding the information being presented [1]. Designing displays that are clear and concise benefit human operators, especially in situations that are time critical or have critical safety issues. Designers can mitigate erroneous behavior that transpired from accidental misinterpretation of information by building elements of forgiveness into the interface. Undo or cancel buttons are examples of good forgiveness implementations; however, these resources may not always be feasible or practical. Systems that produce emergent behaviors, such as collectives, have a low probability of returning to an original state, even when forgiveness elements are incorporated into the design.

2.2.2.4 Memory Principles

Memory is the storage of information and is used to recall knowledge that allows human operators to decide what actions to take [163]. Short-term memory, which encompasses working memory, is a temporary and attention demanding form of memory that examines, evaluates, and transforms information [157]. Three core components constitute working memory: visuo-spatial sketch pad, central executive, and the phonological loop [166]. The primary role of the visuo-spatial sketch pad is to hold and manipulate representations, while the phonological loop is concerned with auditory representations. The central executive allocates resources to the sketch pad or the loop and directs the flow of information. Long term memory stores information for later recall.

SAT Level 1 Replacing memory with visual information knowledge on a display can help alleviate expended effort on behalf of the human operator when interacting with the system [150]. Using standardized words, symbols, and providing checklists can require less memory processing. The designer must consider the quantity of information, even if it is presented more concisely, so as to not overburden the human operator.

SAT Level 2 Designing interfaces to use consistent representations, such as color coding and symbology, to represent particular information can help mitigate encoding the information into long term memory inaccurately [150]. Repeated exposures will restructure the human operator's mental model by overriding the incorrect encoding as well as move working memory into long term memory. Consistent feedback to the human operator will aid in comprehension of the feedback and expedite necessary actions.

SAT Level 3 Human operators are ineffective at predicting future events, because the memory process required to compute every possible outcome is expansive and time consuming. The current system state is considered initially, and simulations of possible future states are generated in order to find and compare instances that appear realistic and likely to occur [150]. Designers can implement systems to incorporate predictive aiding, which mitigates process time, potentially offers more accurate suggestions or predictions, as well as provides projections that are not susceptible to operator bias.

2.2.3 Training

Training is used to prepare human operators for various scenarios that may arise when interacting with known or unknown systems to handle abnormalities, operating under variable and off-nominal conditions, as well as becoming acquainted with new features in the system. Trainers teach human operators how to identify, analyze, and execute appropriate actions. Training is intended to support human operators, help prevent errors, simplify tasks, and promote active learning by providing feedback and imposing practice. Human organizations concern themselves with providing the best training program, delivered in the shortest amount of time for the least amount of money, that leads to the longest retention of knowledge and skill in their human operators [150].

2.2.3.1 Support and Error Prevention

Training can help reduce intrinsic load, which describes the mental workload imposed on human operators when learning how to use and interact with a system [157]. Trainers initially support and guide trainees through a process where the human operator

learns about the system step-by-step to build confidence. The trainer will gradually withdraw from the process in order to encourage learning. Human operators learn how to recognize issues, identify when those issues occur, and more importantly learn how to mitigate errors or diffuse situations quickly in order to maintain safety.

SAT Level 1 Training can be used to evaluate the trustworthiness of a system based on transparency design aspects [110]; however, the results will be impacted by the original level of trust the human operator has in the system. Initial negative perceptions will impact the ability of the training to effectively instruct human operators to develop the necessary skills when interacting with the system [84]. Consistent training can calibrate appropriate trust levels, reduce initial biases, provide knowledge of system capabilities, and help develop a risk assessment of system behavior [84]. The results of training will enable human operators to perceive erroneous state information or inconsistent behaviors from the accumulated knowledge gained with repeated exposures to the system.

SAT Level 2 Different training strategies must be considered when working with decision support systems that provide informative feedback to human operators. The quality and quantity of help and guidance can impact a human operator's trust in the system [112]. Training the operator to develop appropriate mental models of the system's reasoning process will assure adequate trust and mitigate misconceptions associated with the system. Additionally, training can mitigate complacent behavior by instructing the human operator when interactions are needed. Training is a supplemental strategy that must not be used to compensate for poor system design [81].

SAT Level 3 Four error management principles that apply to training situations with complex systems are: 1) encouraging trainees to develop their own mental models of risk assessment and aversion, instead of inflicting one approach, 2) allowing the human operator to make errors in order to learn how to recover from them, 3) introducing heuristics that change the attitudes of trainees, and 4) offering training at various expertise levels [167]. Training will affirm the limitations and uncertainties associated with the system. Human operators can practice implementing various strategies under variable conditions in order to determine how the limitations and uncertainties of the system influence future behavior. Providing projections of the human operator's current actions will help identify which strategies are more applicable for specific scenarios.

2.2.3.2 Task Simplification

Repeated and consistent practice with systems will help simplify complex tasks and develop automatic responses. The trainer can help establish appropriate methods and prioritize steps needed to mitigate issues that arise. Supplemental documentation, such as checklists, can simplify the task for the human operators. Dependency on documentation for more simple tasks will decrease with increased system exposure, as the knowledge gained is stored in long term memory. Jobs that require shift work may benefit from training human operators what techniques work best for smooth transitions that cause little interruption to the system state.

SAT Level 1 Training can be implemented to calibrate appropriate levels of expectancy [167]. Initially, the trainer can force low system complexity and slowly introduce more complex behavior as the human operator becomes more acquainted with the system

and gains an understanding of how the system works. The human operator, through practice with the system, will be able to determine what pertinent information to look for and where to look for that particular information in order to fulfill task objectives. Human-machine interactions will be quicker and produce more effective outputs.

SAT Level 2 Providing feedback to the human operator improves trust calibration and understanding of the system's reasoning process. The quantity of information produced by the system typically increases with system complexity. Designers will be challenged to determine how much information can be supplied to the human operator before feeling overloaded and what information is most crucial. Training can offset the necessity of providing some information, such as how to use an interface, and educate the human operator about the reasoning process before completing real world tasks. The human operator will have a better mental model of what the system is doing, why it is behaving in a particular way, and the reliability of the actions it is taking [110].

SAT Level 3 Repeated interactions with a system, via training, will help human operators identify and understand what potential future projections may emerge from current actions. Training will help human operators determine when to anticipate looking for particular information in order to fulfill task objectives. The human operator can create short cut strategies that yield particular system behavior, such as making decisions quicker or decreasing the number of interactions needed to fulfill an objective. Future projections of the system state and the associated actions required to attain that state can be provided to the human operator. Repeated exposures will increase the efficiency of actions taken by the human-machine system, since the complexity of the task will not appear as complex before training occurred.

2.2.3.3 Active Learning

Practice with a system under variable conditions and tasks can supplement human operators' exposure to infrequent behavior and promote learning, which may have been difficult to accomplish without training [81]. Reoccurring training may help maintain skills necessary to avoid misses or false alarms, as is defined in Signal Detection Theory. Trainers can provide feedback to the human operators regarding common mistakes and provide strategies to mitigate undesired outcomes. Overlearning, beyond the sole intention of mitigating errors, has been shown to improve speed of performance and decrease the rate of forgetting [150]. Developing a training program for highly complex situations is difficult and may require implementing other methods, such as instructional scaffolding, which provides support to foster learning from scenarios [110].

SAT Level 1 Automaticity, the ability to execute tasks that become an automatic response or habit, is the result of learning, repetition, and practice [167]. Training and protocols can enable attention focus [168], aiding in the development of automaticity; however, designers must consider how much automaticity is appropriate, particularly when human-machine systems are executing multiple tasks. Time-sharing skills that teach human operators different resource-allocation strategies are essential if attention must be flexible to various tasks [167]. Designers must consider developing training programs that incorporate the speed and efficiency of automaticity to expedite human operator actions, as well as the attentional flexibility of time-sharing skills, in order for the human operator to not miss important information relative to their tasks. Practice learning both skills with the system will aid human operators.

SAT Level 2 Bjork has described the concept of effort as the human operator's perception of their own memory capabilities [169], a phenomenon that has considerable implications for learning and training. Human operators often believe that they have learned information better than they have. Trainers must monitor the human operator's actions in order to correct erroneous behaviors and guide the human operator towards mission success. Feedback can be provided either during a practice scenario or at the completion of the scenario. Humans cannot recall specific information in great detail over long durations of time; therefore, the trainer must consider how long a practice scenario takes in order to determine when to provide necessary feedback.

SAT Level 3 Error-prevention learning techniques that guide human operators step by step may lead to effective performance and low amounts of effort, because the trainer is instructing the human operators on what to do [167]; however, new types of errors are likely to occur that were not anticipated previously [81]. Training human operators to think strategically through problems via instructional scaffolding, although more time consuming, may be more advantageous, because better explanations can be provided about complex relationships or system limitations and uncertainties [110]. The human operator can develop strategies to understand implications of uncommon future predictions provided by the system and may adapt appropriate actions by using robust strategies. Working with human-machine systems that produce emergent behavior will require human operators to utilize a skill set that can be used across various conditions.

Designing human-machine systems for transparency requires the designer to provide specific system features that are beneficial to the human-machine team, implementation of specific design guidelines, and training the human-machine system under variable conditions. Transparency is an essential principle to ensure that interactions

between the human operator, machine, and environment produce desired outcomes. Transparency will impact the effectiveness of the design. Understanding how the design methods can mitigate challenges in human-machine systems can inform design choices for human-collective systems.

Chapter 3: Experimental Analysis

Two human-collective systems were evaluated on the basis of the direct and indirect transparency factors. The four criteria from Chapter 2.2 were designed into the visualizations in order to provide operators information to complete the desired task. A user evaluation of an abstract human-collective interface system was conducted by Cody *et al.'s* [57], discussed in Chapter 2.0.3.2, while this dissertation conducted a second user evaluation of the same underlying system, but using an interface that visualized each member of a collective. The results from both user evaluations were analyzed as part of the dissertation in order to identify which metrics were most useful in achieving transparency. The analyses contributed towards the effort of developing transparency metrics for analyzing and design guidelines for future human-collective systems.

3.1 Human-Collective Task

The human-collective task involved a single human operator who supervised and assisted four robotic collectives that performed a sequential best-of-n decision-making task, where the human-collective team chose the best option from a finite set of n options [8]. The human-collective team performed two sequential decisions per collective (i.e., moved the collective to a new hub site two consecutive times). The decision-making task entailed the identification and selection of the highest valued target within a constrained 500 m range of the current hub, the collective hub moved to the selected target, and initiated the second target selection decision, which followed the same iden-

tification, selection, and move procedure. The consensus decision-making task required a quorum detection mechanism to estimate when the highest valued target was identified by 30% of the collective [29]. Each collective of 200 simulated Unmanned Aerial Vehicles searched an urban area of approximately 2 km^2.

The four collective hubs were visible at the start of each trial. Targets became visible as each was discovered by a collective's entities. The target's value was assessed by the collectives' entities, who returned to their respective hub to report the target location and value. The collectives were only allowed to discover and occupy targets within their search range, but some targets were within proximity of multiple hubs. A collective's designated search area changed after it moved to a new target to establish the new hub site. The operator was instructed to prevent multiple collectives merging by not permitting their respective hubs to move to the same target. When a collective moved to a target, the hub moved to the target location, and the target was no longer visible to the operator or available to other collectives. The collective that moved its hub to a target's location first, when two collectives were investigating the same target, moved to the target location, while the second collective returned to its previous hub location. Both collectives made a decision when a merge occurred, even though only one collective moved its hub to the respective target location.

3.2 Interface Environment

The general interface design requirements, related to autonomy, control, and transparency, are: 1) enable the operator to estimate the collectives' decision-making process, 2) identify appropriate control mechanisms to influence the decision-making process, and 3) implement the desired control mechanisms [29]. Two models were used. A se-

quential best-of-n decision-making model (M_2) adapted an existing model (M_1), which based decisions on the target's quality (i.e., value) [170]. Information exchanges between a collective's entities was restricted inside the hub, to mimic honeybees. Episodic queuing cleared messages when the individual collective entities transitioned to different states, which resulted in more successful and faster decision completion. Interaction delay and interaction frequency were added as bias reduction methods in order to consider a target's distance from the collective hub and increase interactions among the collectives' entities regarding possible hub site locations. Interaction delay improved the success of choosing the ground truth best targets (i.e., highest value target), and interaction frequency improved decision time. The baseline model (M_3) allowed the individual collective entities to search and investigate potential targets, but was unable to build consensus. The operator was required to influence the consensus-building element and select that final target, based on the consensus. Simulations were ran without an operator for the M_2 model in order to understand the operator's influence on collective behavior, referred to as M_{2SIM}. The M_3 model required operator influence in order to perform the decision-making task; thus, simulation only analysis was not conducted.

The interface control mechanisms allowed the operator to alter the collectives' internal states, including their levels of autonomy, throughout the sequential best-of-n selection process. The collective's entities were in one of four states. *Uncommitted* entities explored the environment searching for targets, and were recruited by other entities while inside of the collective's hub. Collective entities that *favored* a target reassessed the target's value periodically, and attempted to recruit other entities within the collective's hub to investigate the specified target. Collective entities were *committed* to a particular target once a quorum of support was detected, or after interacting with another committed entity. *Executing* collective entities moved from the collective's current hub

location to the selected target's location. A collective operated at a high level of autonomy by executing actions associated with potential targets independently. The operator was able to influence the collective's actions in order to aid better decision-making, effectively lowering the level of autonomy. Communication from the operator with the collective's entities occurred inside the hub in order to simulate limited real-world communication capabilities. The control mechanisms, for influencing the collective, were communicated to the specified hub. Two visualizations were designed and evaluated in order to determine which visualization provided better transparency by facilitating the operator's perception of the collectives' states, comprehension of the collectives' decision-making processes, and means to influence future collectives' actions.

3.2.1 Individual Agents Interface

The Individual Agents (IA) interface, see Figure 3.1, exemplifies a traditional collective visualization by displaying the location of all the individual collective entities [171]. The interface was divided into three primary areas: 1) the central map, 2) the collective request area, and 3) the monitor area. The map, located at the center of the interface visualizes the respective hubs, their individual entities, discovered targets, and other associated information. Both the collectives and targets were rectangular boxes with distinguishing identifiers located at the center of the icon. The collectives had Roman numeral identifiers (I-IV), while the targets used integers (0-15). Discovered targets initially were white and transitioned to a green color when at least two individual collective entities evaluated the target. The highest valued targets were a bright opaque green (e.g., Target 0 in Figure 3.1), while lower valued targets had a more translucent green color (e.g., Target 9 in Figure 3.1). Targets that were within the collective's 500

m search range had different colored outlines, depending on the collective's state: explored targets that were not currently favored had yellow outlines, explored targets that were favored had white outlines (e.g., Target 12 in Figure 3.1), and targets that were abandoned have red outlines (e.g., Target 13 in Figure 3.1).

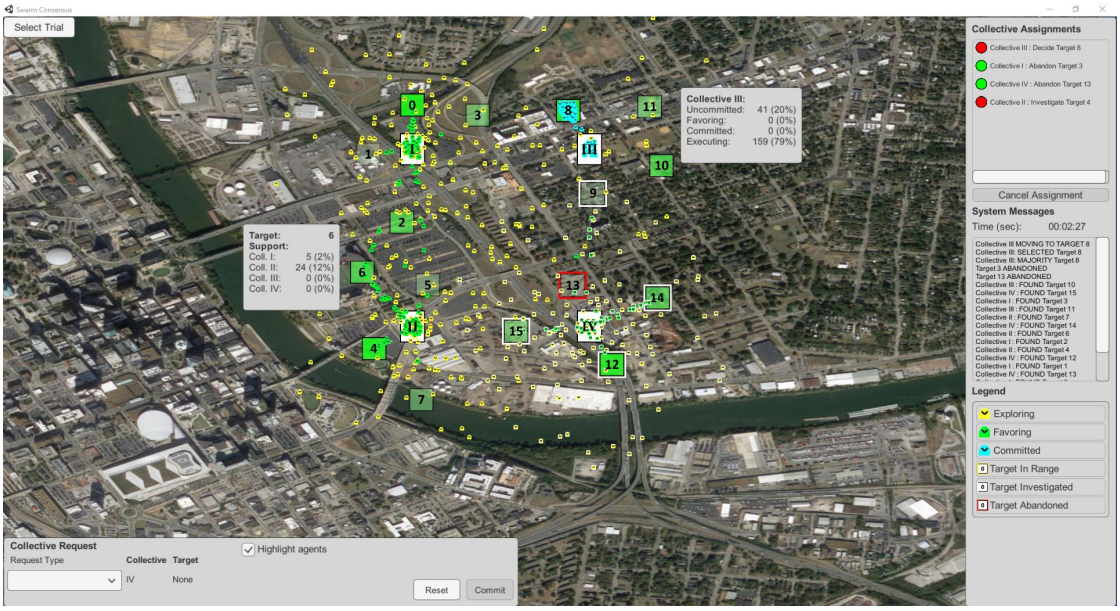


Figure 3.1: The Individual Agents (IA) interface two and half minutes into a trail, showing four collectives (rectangles with Roman numerals), and the sixteen discovered targets (rectangles with integers). The target's value is represented by the green color, where higher values were brighter. The legend in the lower right corner identifies the individual collective entity state information and target range information.

The individual collective entities began each trial by exploring the environment in an uncommitted state, which transitioned to favoring as targets were assessed and supported. The individual collective entities committed to a target once 30% of the collective (60 individual entities) favored a particular target. The collective executed a move to the selected target's location once 50% of the collective (100 entities) favored the target. The individual collective entities' state information was conveyed via indi-

vidual collective entity color coding: uncommitted (yellow), favoring (green), committed (blue), and executing (blue). A legend of the collective entities' and target border colors was provided in the lower right-hand corner, see Figure 3.1. The number of individual collective entities in a particular state, or supporting a target was provided via the collective hub and target information pop-up windows, which provided detailed information, represented as gray rectangular boxes, displayed directly on the map in Figure 3.1. The information pop-up windows, when accessed, appeared in a particular location relative to the respective collective's hub or target. The operator was able to move the information windows by dragging the pop-up window to a desired location.

The operator had the ability to influence an individual collectives' current state via the collective request area, located on the lower left-hand side of Figure 3.1. The *investigate* command permitted increasing a collective's support for an operator specific target. Ten uncommitted entities (5% of the collective population) transitioned to the favoring state after receiving and acknowledging the investigate command. Additional support for the same target was achieved by reissuing the investigate command repeatedly. The *abandon* command reduced a collective's support for a specific target by transitioning favoring individual entities to the uncommitted state. The abandon command only needed to be issued once in order for the collective to ignore a specified target. A collective's entities stopped exploring alternative targets and moved to the operator selected target when the *decide* command was issued, which was a valid request when at least 30% of the collective supported the operator specified target. An operator using the IA interface was no longer able to further influence a collective once the decide command was issued. The process to issue a command first required the selection of the desired command from the drop down menu, then selection of the desired collective and target, and the request was completed by clicking on the commit button. The reset button,

cleared entered information, allowing the operator to select new request information. The highlight agents selection box identified which individual entities belonged to a particular collective. When the highlight agents box was selected, the specific individual collective entities associated with a hub were highlighted with a white border to distinguish them from the other entities. The highlight was deactivated by deselecting the highlight agents selection box, which removed the check mark.

The collective assignments area logged the operator's issued commands, shown in the upper right-hand corner of the monitor area in Figure 3.1. The log displayed what commands were issued with respect to particular collectives and targets (e.g., Collective I: Abandon Target 3). The green and red circles next to each command signified whether the command was completed (red) or currently active (green). An investigate command initially had a green circle and transitioned to red once ten individual entities received and acknowledged the investigate command for a particular target. Issued abandon commands for a particular collective and target remained active (constant green circle). Once a collective reached a decision, all prior commands associated with that particular collective were removed from the collective assignments log. The only command the operator was able to cancel was the abandon command, which required selecting the desired abandon command text line, in Figure 3.1, the "Collective I: Abandon Target 3", and then selecting the cancel assignment button.

System messages indicated the operator and collectives' actions. The illegal message was displayed when an operator requested an invalid command, and explained why the requested action was not viable. Three situations resulted in illegal messages. The first arose when the operator attempted to issue an investigate command for targets that were outside of the collective's search region. The second situation occurred when the operator attempted to abandon newly discovered targets that did not have an assigned

value (white targets). The last situation arose when the operator attempted to issue decide commands when less than 30% of the collective supported a target.

3.2.2 Collective Interface

The Collective interface [29], shown in Figure 3.2, provides an abstract visualization that does not present individual collectives' entities. The Collective interface was divided into the same three primary areas as the IA interface: 1) the central map, 2) the collective request area, and 3) the monitor area. The operator commands were and func-

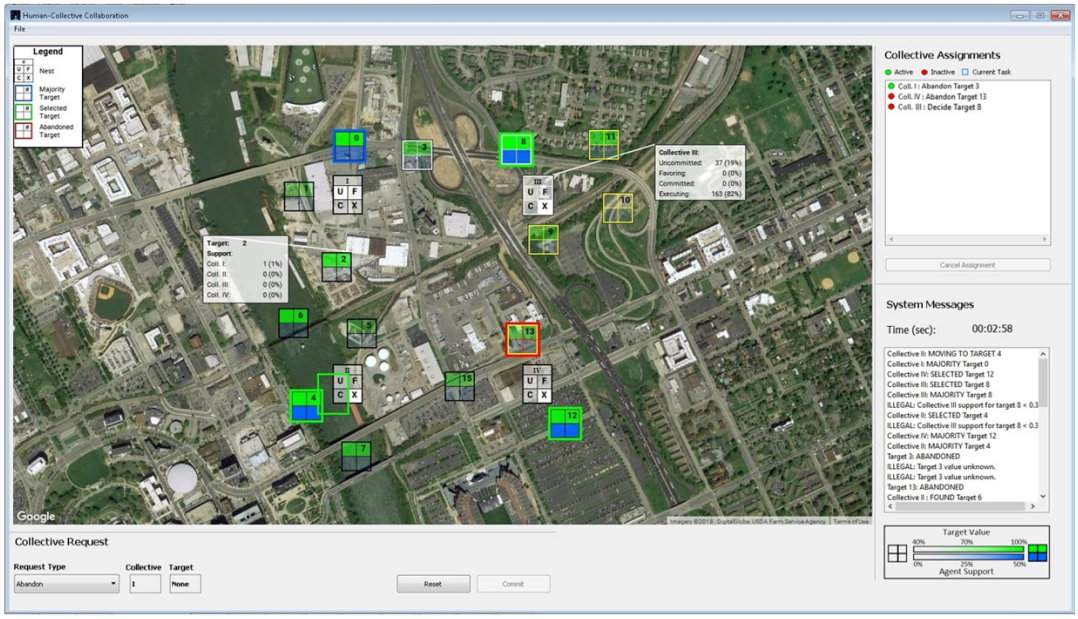


Figure 3.2: The Collective interface mid-way through a trail scenario, showing the current locations of the four collectives (rectangles with Roman numerals) and the locations of the discovered targets (green and blue squares with integer identifiers). The top half of each target indicates the target's relative value (green) and the bottom half indicates the support of the highest supporting collective (blue). The legend in the upper left hand corner identifies the target range information.

tioned the same as those in the IA interface. The collectives were represented as gray and white rectangles with four quadrants and Roman numeral identifiers located at the top center of the icon. The collective state information was conveyed via the collective hub icon's quadrants, color coding, and information pop-up windows. The collective icon's contained four state quadrants (uncommitted (U), feedback (F), committed (C), and executing (X)), which represented the number of individual entities in each state, where a brighter white quadrant equated to larger numbers of individual collective entities. The square target icons had integer identifiers positioned on the upper right hand corner. Target icons contained two sections: 1) the top-half green section represented the target's value, where the brighter and more opaque the green, the higher the value (e.g., Target 8 in Figure 3.2), and 2) the bottom-half blue section indicated the number of individual entities favoring a particular target, where the brighter and more opaque the blue, the higher the number of collective entities (e.g., Target 12 in Figure 3.2).

The collective interface operated similarly to the IA interface with some distinctions. A target was outlined in blue, demonstrated by Target 0 in Figure 3.2, when the collective's support exceeded 30%. The target transitioned to a green outline, and the collective was outlined in green when the collective began executing a move to the target's location. The collective's outline moved from the hub to the target's location to indicate the hub's transition to the selected target. Once the collective's outline reached the selected target location the hub appeared at that location. The interface's legend appeared in the upper left corner, see Figure 3.2.

3.3 IA and Collective User Evaluations Experimental Design

The experimental design for the two single operator-collective evaluations are discussed. The associated independent variables, the experimental procedure, and the IA and Collective evaluation participants are described in detail.

3.3.1 Independent Variables

The independent variables for the single operator-collective evaluations were the within model variable (M_1, M_2, and M_3) and the trial difficulty (overall, easy, and hard). The IA evaluation excluded the M_1 model, because the assessment was interested understanding the differences between a more advanced best-of-n model (M_2) versus a baseline model (M_3). Trials that had a larger number of high valued targets in closer proximity to a collective's hub were deemed *easy*, while *hard* trials placed high valued targets further away from the collective's hub. The independent variables associated with each evaluation are identified in Table 3.1.

Table 3.1: Independent variables associated with single operator-collective evaluations.

Independent Variable	**IA Evaluation**	**Collective Evaluation**
Model	M_2, M_3	M_1, M_2, M_3
Decision Difficulty	All	All

3.3.2 Experimental Procedure

The experimental procedure for both user evaluations required participants to complete a demographic questionnaire (Appendix A), and a Mental Rotations test [172]. The IA participants also completed a Working Memory Capacity assessment. Upon comple-

tion of the demographic data collection, participants received training and practiced using their interface. Practice sessions occurred prior to each trial in order to ensure familiarity with the underlying models. The M_2 model trial was always completed first in the IA evaluation, in order to alleviate any learning effects from using the M_3 model. The collective evaluation randomized the order of the M_1 and M_2 models, which were always presented before the M_3 model.

The participants were instructed that the objective was to aid each collective in selecting and moving to the highest valued target two sequential times. A trial began once the practice session was completed. Each trial was divided into two components (one easy and one hard) of approximately ten minutes each. Splitting each trial into two components allowed the environment to reset with 16 new (not initially visible) targets. The easy trial contained higher valued targets close to the hub, while the hard trial placed high valued targets further away. The easy and hard trial orderings were randomly assigned, and counterbalanced across the participants. The situational awareness (SA) probe questions [29] (Appendix B), were intended to serve as a secondary task and were asked beginning at 50 seconds into the trial and repeated at one-minute increments. Six SA probe questions were asked during each trial component, resulting in twelve total SA probe questions per trial. The trial was terminated once the team completed eight decisions, two per collective, or once six decisions were made, if the trial length exceeded the ten-minute limit. Decision times were not limited. A post-trial questionnaire (Appendix C) was completed after each trial and the post-experiment questionnaire (Appendix D) was completed before the evaluation termination.

3.3.3 Participants

Prior to enrollment in the evaluation, potential participants were screened for color blindness. Individuals who self-identified as color blind were excluded from the evaluations. The participants from both evaluations who were successfully enrolled completed a demographic questionnaire, which collected information regarding age, gender, education level, *weekly hours on a desktop or laptop* (0, less than 3, 3-8, and more than 8), and their *video game proficiency* from little to no proficiency (1) to high proficiency (7). The *Mental Rotation Assessment* [172] required participants to judge three-dimensional object orientation to assess spatial reasoning within a scoring range of 0 (low) to 24 (high). The mode is reported in parenthesis for questions that required selection to a group. The additional *Working Memory Capacity* assessment, which was only completed by participants from the IA user evaluation, evaluated the participant's performance of higher-order cognitive tasks [173]. A reading span test required participants to determine whether a sentence was accurate while recalling a series of letters interspersed between sentences. Letter recall accuracy was measured as the proportion of correctly recalled letters to the total number of letters.

3.3.3.1 IA Evaluation Participants

Fourteen females and nineteen males completed the IA evaluation at Oregon State University. The main (25) age range was 18 to 30 years, with seven participants between 31 and 50, and one was 60 and older. Many participants were in the process of obtaining (8) or had an undergraduate degree (13), a master's degree (9), or a doctorate degree (1). The mean weekly hours on a desktop or laptop was 3.79, with a standard devia-

tion (SD) = 0.5, median = 4, minimum (min) = 2, and maximum (max) = 4. The video game proficiency ranking mean was 4.61 (SD = 1.93, median = 5, min = 1, max = 7). The Mental Rotation Assessment [172] mean was 12.36 (SD = 5.85, median = 12, min = 3, and max = 24) [171]. The Working Memory Capacity sentence accuracy mean was 86.14 (SD = 9.73, median = 89.5, min = 59, and max = 98) and letter recall mean was 74.07 (SD = 14.79, median = 78, min = 43, and max = 94) [171]. Five participants were excluded from the analysis due to inconsistent methodology (1) and software failure (4).

3.3.3.2 Collective Evaluation Participants

Twenty-eight participants, fifteen females and thirteen males, from Vanderbilt University, completed the Collective evaluation. The majority of participants (24) were between 18 and 30 years old, with four between 31 and 50. Most of the participants completed high school and were in the process of obtaining (11) or had completed an undergraduate degree (13). The weekly hours participant's used a desktop or laptop was slightly higher than that of the IA (mean = 3.86, SD = 0.45, median = 4, min = 2, and max = 4). Video game proficiency was ranked lower than the IA (mean = 3.61, SD = 2.23, median = 2.5, min = 1, and max = 7). The participants' Mental Rotations Assessment scores were also slightly lower than the IA evaluation (mean = 10.93, SD = 5.58, median = 10, min = 1, and max = 24) [171].

3.4 Analyses of Transparency Experimental Design

Two distinct analyses of transparency were conducted, the Visualization Analysis and the Model with Visualization Analysis, using the results from the IA and Collective

user evaluations. The research questions and independent variables associated with the transparency analyses experimental design are discussed. The description of the dependent variables are associated with the respective research questions in Chapter 4.

3.4.1 Research Questions

The primary between-visualization analysis (Visualization Analysis) research question was to determine *which visualization achieved better transparency?* Four secondary questions were developed in order to investigate how the visualization impacted a direct transparency factor, exclusive of trust. The first research question (R_1) focused on understanding *how the visualization influenced the operator*. Individual differences, such as experience level, will impact an operator's ability to interact with the visualization and may cause different responses (e.g., loss of situational awareness or more workload). A visualization that can aid operators with different capabilities is desired. The explainability factor was encompassed as R_2, which explored whether *the visualization promoted operator comprehension*. Perception and comprehension of the visualized information are necessary to inform future actions. Understanding *which visualization promoted better usability*, R_3, will aid designers in promoting effective transparency in human-collective systems. *Which visualization promoted better human-collective performance* was also assessed (R_4). A system that performs a task quickly, safely, and successfully is ideal.

The primary and secondary research questions for the within-model and between-visualization analysis (Model with Visualization Analysis) expanded on research questions R_1 - R_4. The primary question was to determine *which model and visualization achieved better transparency?* Research question R_5 focused on understanding *how the model and visualization influenced the operator*. The explainability factor was encompassed

in R_6, which explored whether *the model and visualization promoted operator comprehension*. Promoting transparency in human-collective systems required understanding *which model and visualization promoted better usability*, R_7. The final research question (R_8) assessed *which model and visualization promoted better human-collective performance*.

3.4.2 Independent Variables

The independent variables associated with the analyses were the between visualization variable, IA versus Collective, the within model variable, M_2 and M_3, and the trial difficulty (overall, easy, and hard). The M_1 model was excluded from both analyses for the same reason mentioned in Chapter 3.3.1, because the assessment was interested understanding the differences between an advanced best-of-n model (M_2) and the baseline model (M_3). *Easy* trials had a larger number of high valued targets in closer proximity to a collective's hub, while *hard* trials placed high valued targets further away from the collective's hub. The independent variables associated with each analysis and the respective research questions are identified in Table 3.2.

Table 3.2: Independent variables associated with respective analyses and research questions.

Independent Variable	**Analysis**	**Research Question**
Visualization	Both	R_1 - R_8
Model	Model with Visualization	R_5 - R_8
Decision Difficulty	Both	R_1 - R_8

Chapter 4: Results and Discussions

The analyses for all of the research questions are based on a total of 56 participants from both the IA and Collective evaluations. The first twelve decisions made per participant using each model were analyzed. The majority of the objective metrics were analyzed by SA level (overall (SA_O), perception (SA_1), comprehension (SA_2), and projection (SA_3)), decision difficulty (overall, easy, and hard), timing with respect to a SA probe question (15 seconds before asking, while being asked, or during response to a SA probe question), or per participant. Non-parametric statistical methods, including Mann-Whitney-Wilcoxon tests with one degree of freedom (DOF = 1) and Spearman Correlations, were calculated due to a lack of normality. The correlations were with respect to SA probe accuracy and selection success rate. The Collective evaluation results [29] were reanalyzed as part of this dissertation using the same methods. Additional metrics that were not presented in this chapter are provided in Appendices E - G.

4.1 Visualization Analysis

The primary objective of the between-visualization analysis was to determine which visualization achieved better transparency. Four secondary research questions were created to assess how transparency influenced individuals with different capabilities, operator comprehension, visualization usability, and human-collective performance. The subset of direct and indirect transparency factors (Figure 2.2) were assessed in the Visualization Analysis and are identified in Figure 4.1. The hypotheses, metrics, results,

and discussions for the between-visualization analysis are presented in Chapters 4.1.1 - 4.1.4, which correspond to research questions R_1 - R_4. A research question specific representation of the analyzed direct and indirect transparency factors is provided. The Visualization Analysis is concluded with a final discussion that incorporates the discussions from each respective secondary research question.

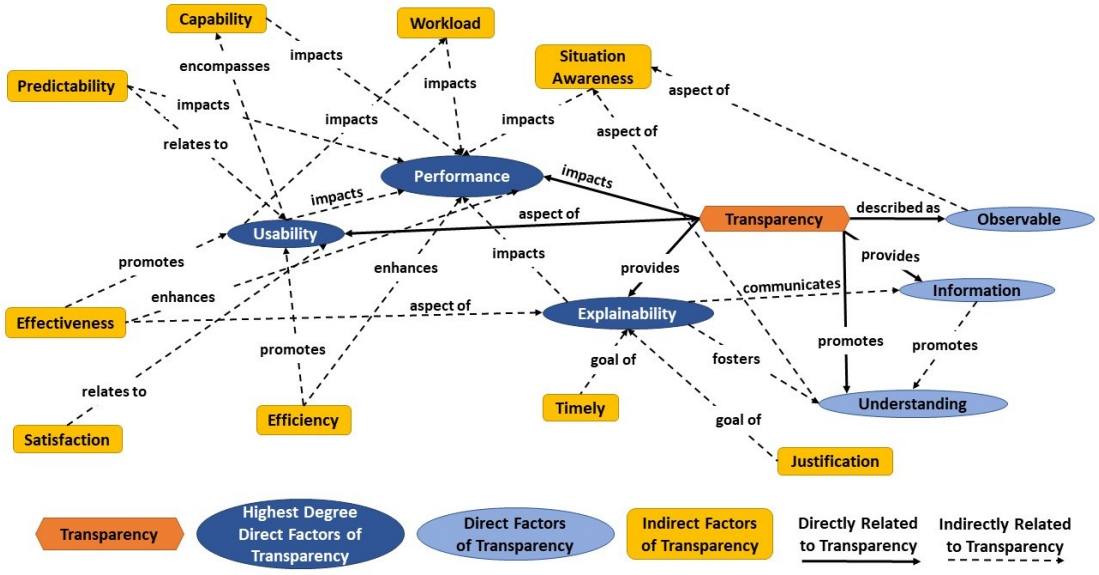


Figure 4.1: The analyzed direct and indirect transparency factors included in the Visualization Analysis.

4.1.1 R_1: Visualization Influence on Human Operator

Understanding *how the visualization influenced the operator*, R_1, is necessary to determine if the transparency embedded into the system design aided operators with different capabilities. The associated objective dependent variables were (1) the operator's ability to influence the collective in order to choose the highest *target value*, (2) SA, (3) visu-

alization clutter, and (4) the operator's spatial reasoning capability (Mental Rotations Assessment). The specific direct and indirect transparency factors related to R_1 are identified in Figure 4.2. The relationship between the variables, the corresponding hypotheses, and the direct and indirect transparency factors, are identified in Table 4.1. Additional relationships (not shown in Figure 2.2) between the variable and the direct or indirect transparency factors are identified due to correlation analyses.

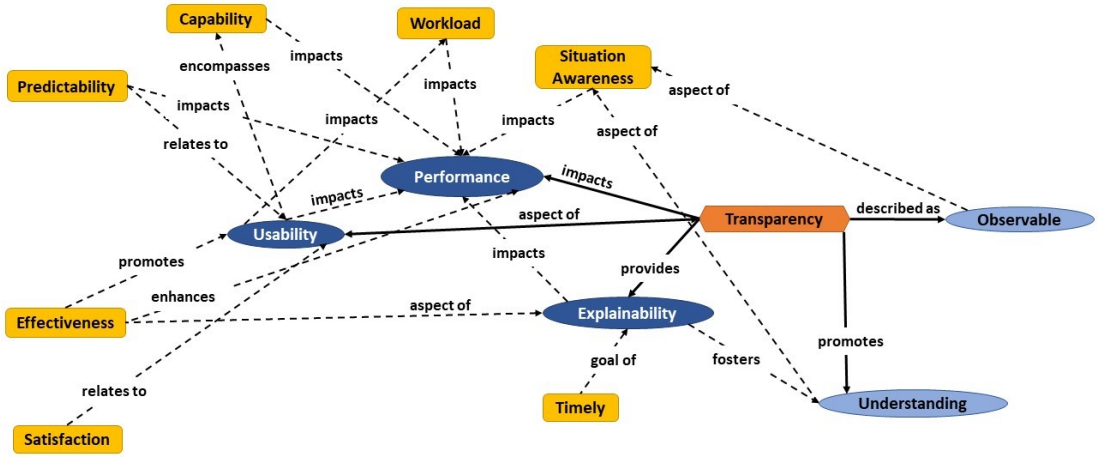


Figure 4.2: R_1 concept map of the assessed direct and indirect transparency factors.

Operators may have performed differently depending on their individual differences. It was hypothesized (H_1) that operators using the Collective visualization will experience significantly higher SA and lower workload. SA represents an operator's ability to perceive and comprehend information in order to project future actions that must be taken in order to fulfill a task [60]. Usability influences the perception of information [2] and will impact workload, which is the amount of stress an operator experiences in order to accomplish a task during a particular duration of time [150]. It was hypothesized (H_2) that operators with different individual capabilities will not perform

Table 4.1: Visualization influence on the human operator objective (obj) and subjective (subj) variables (vars), relationship to the hypotheses (H), as well as the associated direct and indirect transparency factors, are presented in Figure 2.2.

		Transparency Factors											
		Direct					**Indirect**						
Obj Vars	**H**	**Explainability**	**Observable**	**Performance**	**Understanding**	**Usability**	**Capability**	**Effectiveness**	**Predictability**	**SA**	**Satisfaction**	**Timing**	**Workload**
Target Value	H_1					✓							
SA Probe Accuracy	H_1	✓	✓	✓	✓			✓	✓	✓			
Local Clutter	H_1	✓	✓	✓	✓	✓		✓	✓	✓			
Global Clutter	H_1	✓	✓	✓	✓	✓		✓	✓	✓			
Mental Rotations Assessment	H_2			✓			✓						
Subj Vars													
Weekly Hours on a Desktop or Laptop	H_2			✓			✓						
Video Game Proficiency	H_2			✓			✓						
NASA Task Load Index	H_1, H_3	✓		✓			✓				✓	✓	✓
3-D Situational Awareness Rating Technique	H_1	✓			✓		✓	✓		✓			

significantly different using the Collective visualization. An ideal visualization will enable operators with different capabilities to perceive, comprehend, and influence collectives relatively the same. Training can alleviate any disparities between operators, but is only intended to supplement the system's design. The operator's attitude and sentiments towards a system, which is dependent on system usability, provides essential information related to the system's design [105]. Good designs promote higher operator

satisfaction. It was hypothesized (H_3) that operators using the Collective visualization will experience significantly less frustration (i.e., higher satisfaction).

4.1.1.1 Metrics and Results

Assessing variables, such as the selected target value for each human-collective decision, is necessary in order to determine whether operators were able to perceive the target value correctly and influence the collectives positively. The objective of the human-collective team was to select the highest valued target for each decision from a range of target values (67 to 100). The selected target value is the average of all target's respective values that were selected by the human-collective teams during a trial. The descriptive statistics for the selected target value per decision difficulty (i.e., overall, easy, and hard) are shown in Table 4.2. Operators using the Collective visualization were able to influence the collective to chose higher valued targets, regardless of decision difficulty, on average; however, the Mann-Whitney-Wilcoxon test identified no significant effects between visualizations for the selected target value.

Table 4.2: Selected target value descriptive statistics by decision difficulty, where the maximum possible value was 100 and the minimum possible value was 67.

	Decision Difficulty	**Mean (SD)**	**Median (Min/Max)**
IA	Overall	90.29 (7.11)	95 (67/97)
	Easy	90.21 (7.29)	95 (67/97)
	Hard	90.4 (6.88)	94 (68/96)
Collective	Overall	92.05 (5.08)	95 (68/96)
	Easy	92.09 (5.54)	95 (68/96)
	Hard	92 (4.5)	95 (78/96)

The SA dependent variable was *SA probe accuracy*, the percentage of correctly answered SA probes questions, which assessed the operator's SA during a trial [29]. Each

question corresponded to the three SA levels: perception, comprehension, and projection [60]. Participants were asked five SA_1, four SA_2, and three SA_3 questions. The SA_1 questions determined the operator's ability to perceive collective and target information, such as "What collectives are investigating Target 3?" The operator's comprehension of information was represented by the SA_2 questions, such as "Which Collective has achieved a majority support for Target 7?" SA_3 questions related to the operator's ability to estimate the collectives' future state, such as "Will support for Target 1 decrease?" An overall SA value, SA_O, represented the percent of correctly answered SA probes out of 12 total. The SA probe accuracy descriptive statistics are shown in Table 4.3 [171]. Operators using the Collective visualization had higher SA probe accuracy. The Mann-Whitney-Wilcoxon tests (n = 56) found highly significant effects between visualizations for SA_O (U = 702, $\rho < 0.001$) and SA_1 (U = 714.5, $\rho < 0.001$). Moderately significant effects were found for SA_2 (U = 572.5, $\rho < 0.01$) and SA_3 (U = 554, $\rho < 0.01$).

Table 4.3: SA probe accuracy (%) descriptive statistics by SA level.

	SA Level	**Mean (SD)**	**Median (Min/Max)**
IA	SA_O	65.3 (18.87)	68.33 (16.67/83.33)
	SA_1	58.57 (23.05)	60 (20/100)
	SA_2	72.32 (21.88)	75 (25/100)
	SA_3	65.48 (34.52)	66.67 (0/100)
Collective	SA_O	89.88 (10.96)	91.67 (58.33/100)
	SA_1	91.67 (11.11)	100 (66.67/100)
	SA_2	88.39 (14.6)	100 (60/100)
	SA_3	89.88 (20.46)	100 (33.33/100)

Local and global clutter percentages were analyzed for each SA probe question. Clutter is defined as the area occupied by objects on a display, relative to the total area of the display [150]. Presenting too much information in close proximity to one another will require the operator to search longer for information [150] and can negatively in-

fluence the accuracy of the SA probe question responses. Area coverage for each 2-D item was calculated by the number of pixels the item covered on the computer visualization. The conversion between meters and pixels was different for each visualization due to differences in the display monitor size and software program. One meter for the IA visualization was approximately 1.97 pixels per meter and the Collective visualization was approximately 2.3 pixels per meter. The *local clutter percentage* variable was the percentage of area obstructed by items that were displayed within the 500 *m* (i.e., approximately 254 pixels for the IA visualization and 218 pixels for the Collective visualization) circular radius from the center of the collective, or target of interest in the SA probe question. Collective IV, for example, is the collective of interest in the following SA probe question: "What is the highest value target available to Collective IV?" The items obstructing the 500 *m* radius when using the IA visualization, in Figure 3.1, for the previous SA probe question are: the Collective IV, Targets 9 and 12-15, and 200 individual entities. Some SA probe questions encompassed more than one collective or target of interest, which required calculating the local clutter percentage for each collective or target and summing the values together. Calculations first required converting meters into pixels in order to ensure equivalent units. The Collective visualization computer display size was unknown; therefore, local and global clutter percentage calculations use the corresponding item and computer display dimensions from the IA visualization. Local clutter was calculated using Equation 4.1:

$$LocalClutter(\%) = \sum \left(\frac{LHA + LHTA + LTA + LAICE + LTIW + LCIW}{\pi \cdot 500^2} \right) \cdot 100, \tag{4.1}$$

where LHA represented the area corresponding to the number of collective hubs (2464 $pixels^2$ per hub) inside the 500 m radius. The area corresponding to the number of highlighted targets (2350 $pixels^2$ per highlighted target), which had outlines and were in range of the selected collective were represented as LHTA, while the targets that were not highlighted (1720 $pixels^2$ per target) were denoted as LTA. LAICE represented the area corresponding to the number of individual collective entities (64 $pixels^2$ per individual entity) inside of the 500 m radius, and was excluded from the Collective visualization local clutter percentage calculation, because no individual entities were displayed. The individual collective entities were confined to the 500 m search radius around their respective collective hub; therefore, the calculation assumes that the 200 entities associated with each collective are inside of the local radius. The area corresponding to the number of target information pop-up windows (32922 $pixels^2$ per target information pop-up window) was represented as LTIW, and the corresponding collective information pop-up windows (25740 $pixels^2$ per collective information pop-up window) were represented as LCIW. Only target or collective information pop-up windows that belong to targets or collectives inside of the 500 m radius were considered. The Collective evaluation did not record which particular collective information pop-up windows were visible; therefore, LCIW was excluded from the local clutter percentage calculation for the Collective visualization.

The local clutter percentage descriptive statistics 15 seconds before asking, while being asked, and during a SA probe response are provided in Table 4.4. The Collective evaluation did not record the SA probe response time; therefore, the average response time per SA level from the IA evaluation was used for all calculations during response to a SA probe question. The maximum local clutter percentage was 177%, which indicated that the area covered by the associated items of the collective or target of interest

Table 4.4: Local clutter percentage descriptive statistics 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	SA Level	Mean (SD)	Median (Min/Max)
IA	Before	SA_O	33.6 (21.66)	26.13 (9/124)
		SA_1	30.79 (19.53)	24.4 (9/122.33)
		SA_2	41.54 (24.39)	37.75 (9/124)
		SA_3	28.61 (19.36)	22.23 (9.08/97.7)
	Asking	SA_O	34.42 (22.16)	28 (9/124)
		SA_1	31.91 (20.54)	25 (9/122)
		SA_2	41.67 (24.74)	37.17 (9/124)
		SA_3	29.73 (19.54)	24.21 (9/97.67)
	Responding	SA_O	34.26 (22.25)	27.5 (8/124)
		SA_1	31.84 (20.61)	24.83 (9/122)
		SA_2	41.28 (24.96)	36.5 (8/124)
		SA_3	29.68 (19.6)	24 (9/98)
Collective	Before	SA_O	35.42 (28.08)	25.44 (9/177)
		SA_1	34.09 (29.09)	24.04 (9/151.9)
		SA_2	35.81 (27.53)	26.86 (10.3/177)
		SA_3	37.98 (26.78)	28.23 (10.38/131.47)
	Asking	SA_O	35.37 (28.78)	25.75 (9/176.5)
		SA_1	34.24 (30.37)	23.63 (9/176.5)
		SA_2	36.47 (26.76)	27.4 (10.2/130.4)
		SA_3	36.35 (28.27)	27.25 (9/147.56)
	Responding	SA_O	35.6 (29.19)	25.8 (9/176.4)
		SA_1	34.29 (29.84)	25 (9/176.4)
		SA_2	36.55 (27.98)	27 (10/130)
		SA_3	37.24 (29.86)	26.57 (9/147.57)

in the SA probe exceeded the 500 *m* radius. Local clutter percentages larger than 100% were attributed to the area covered by the collective and target information pop-up windows. The location of the information pop-up windows were not recorded; therefore, the maximum area coverage was considered when information pop-up windows did not occlude items in the environment. The maximum area coverage condition was not reflective of the real trial environment, where information pop-up windows covered

items on the central map. The IA visualization had lower local clutter percentage, regardless of when the metric was collected for SA_O, SA_1, and SA_3. No correlations were found between local clutter percentage and SA probe accuracy.

The *global clutter percentage*, calculated using Equation 4.2, was the percentage of area obstructed by all objects displayed on the entire IA computer display (2073600 $pixels^2$), since the Collective computer display was unknown.

$$GlobalClutter(\%) = \left(\frac{ICA + GHA + GHTA + GTA + GAICE + GTIW + GCIW}{2073600} \right) \cdot 100, \tag{4.2}$$

where ICA represented the area of the static interface components (493414 $pixels^2$), which encompassed the program bar, the Microsoft Windows program bar, the select trial button, the collective request area, and the monitor area. GHA represented the area covered by Collective hubs I-IV (9856 $pixels^2$), which were visible throughout the duration of a trial. The area corresponding to the number of highlighted targets (2350 $pixels^2$ per highlighted target), which had outlines and were in range of the currently selected collective were represented as GHTA. Remaining targets that were not highlighted (1720 $pixels^2$ per target), were represented as GTA. GAICE represented the area encompassed by 800 individual collective entities (51200 $pixels^2$), which was only considered for the IA visualization. The area corresponding to the number of target information pop-up windows (32922 $pixels^2$ per target information pop-up window) was represented as GTIW and the corresponding collective information pop-up windows was represented as GCIW (25740 $pixels^2$ per collective information pop-up window).

The global clutter percentage descriptive statistics 15 seconds before asking, while being asked, and during response to a SA probe question are shown in Table 4.5. The

Table 4.5: Global clutter percentage descriptive statistics 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	SA Level	Mean (SD)	Median (Min/Max)
IA	Before	SA_O	30.2 (3.06)	28.83 (27/40.22)
		SA_1	29.88 (2.8)	28.5 (27/40.22)
		SA_2	30.41 (3.05)	29.1 (27/40)
		SA_3	30.45 (3.45)	28.85 (27/40)
	Asking	SA_O	30.25 (3.13)	29 (27/40)
		SA_1	29.95 (2.91)	28.58 (27/40)
		SA_2	30.41 (3.12)	29 (27/40)
		SA_3	30.52 (3.49)	29 (27/40)
	Responding	SA_O	30.09 (3.02)	29 (27/40)
		SA_1	29.83 (2.81)	28.5 (27/40)
		SA_2	30.22 (3)	29 (27/40)
		SA_3	30.37 (3.38)	28.79 (27/40)
Collective	Before	SA_O	31.37 (4.97)	29.21 (27.88/53)
		SA_1	31.38 (5)	29.13 (28/52.4)
		SA_2	31.25 (5.09)	29.21 (27.88/53)
		SA_3	31.56 (4.76)	29.54 (28/52)
	Asking	SA_O	31.43 (5.13)	29.17 (28/53)
		SA_1	31.24 (5.26)	29 (28/53)
		SA_2	31.52 (5.2)	29.22 (28/51.8)
		SA_3	31.69 (4.78)	29.75 (28/48)
	Responding	SA_O	31.41 (5.15)	29 (28/53)
		SA_1	31.43 (5.43)	29 (28/53)
		SA_2	31.34 (5.08)	29 (28/51.5)
		SA_3	31.49 (4.66)	29.31 (28/48)

IA visualization had a lower global clutter percentage, regardless of when the metric was assessed across all SA levels. The Mann-Whitney-Wilcoxon tests found highly significant effects between visualizations when responding to a SA probe question (n = 670) for SA_O (U = 64442, $\rho < 0.001$). Moderate significant effects were found for SA_O 15 seconds before asking (U = 64188, $\rho < 0.01$) and while being asked a SA probe question (U = 63728, $\rho < 0.01$). Significant effects were found 15 seconds before asking a SA

probe question for SA_1 (n = 294, U = 12487, ρ = 0.02) and SA_3 (n = 152, U = 3445.5, ρ = 0.03); while being asked a SA probe question for SA_1 (U = 12301, ρ = 0.03) and SA_3 (U = 3452, ρ = 0.05); and during the response to a SA probe question for SA_1 (U = 12216, ρ = 0.04). The Spearman correlation analysis revealed a weak correlation for the Collective visualization for SA_1 between global clutter percentage 15 seconds before asking a SA probe question and SA probe accuracy (r = 0.16, ρ = 0.05).

The Mental Rotations Assessment [172], which assessed the operator's spatial reasoning, identified no significant effects between visualizations. A Spearman correlation analysis revealed weak correlations between the Mental Rotations Assessment and SA probe accuracy when using the IA visualization for SA_O (r = 0.17, $\rho < 0.01$), SA_1 (r = 0.18, ρ = 0.03), and SA_2 (r = 0.27, $\rho < 0.01$). The Mann-Whitney-Wilcoxon tests identified no significant effects between visualizations for the weekly hours spent using a desktop or laptop and video game proficiency. Weak correlations were found between weekly hours using a desktop or laptop and SA probe accuracy for the IA visualization for SA_O (r = 0.12, ρ = 0.04) and SA_1 (r = 0.21, ρ = 0.01), as well as when using the Collective visualization for SA_2 (r = 0.21, ρ = 0.02). No correlations were found between video game proficiency and SA probe accuracy.

The NASA Task Load Index (NASA-TLX) assessed the six workload subscales and the weighted overall workload [174]. The descriptive statistics for the *NASA-TLX* demands imposed on the operator are presented in Table 4.6. The Collective visualization imposed a lower overall workload, had lower physical and temporal demands, and caused less frustration [57]. The IA visualization imposed a lower mental demand, which had a significant effect between visualizations (n = 56, U = 515, ρ = 0.04) and less effort. The IA visualization had a higher performance with a highly significant effect between visualizations (U = 159.5, $\rho < 0.001$).

Table 4.6: NASA-TLX descriptive statistics.

	Overall and Subscales	Mean (SD)	Median (Min/Max)
IA	Overall	62.14 (14.81)	65.67 (24/85.67)
	Mental	19.25 (8.8)	20 (0/33.33)
	Physical	1.68 (3.32)	0 (0/13)
	Temporal	11.75 (8.24)	9.67 (0/28.33)
	Performance	10.69 (5.87)	8.83 (2.67/25)
	Effort	11.35 (6.68)	11 (2.67/28.33)
	Frustration	7.43 (8.36)	4.67 (0/33.33)
Collective	Overall	57.06 (16.47)	56.83 (5.67/83.33)
	Mental	23.58 (6.34)	25 (3/31.67)
	Physical	0.46 (1.17)	0 (0/4.67)
	Temporal	10.94 (7.67)	10.33 (0/24)
	Performance	5.1 (4.7)	3.67 (0/21.33)
	Effort	12.32 (6.26)	13 (2/25.33)
	Frustration	4.65 (6.84)	1.83 (0/30)

The 3-D Situational Awareness Rating Technique (SART) measured the operator's perceived situational understanding, demand on attentional resources, and supply of attentional resources [175]. An overall score was calculated using the standard calculation. The *SART* descriptive statistics are shown in Table 4.7 [57, 171]. The minimum SART score was -1, which was unexpected as a negative score requires the supply of attentional resources to exceed the demand on attentional resources and a low perceived situational understanding. Both of these conditions are highly unlikely. The Collective visualization had a higher overall score, more situational understanding, high demands of attentional resources (although nearly the same as the IA visualization), and more supply of attentional resources, compared to the IA visualization. The Mann-Whitney-Wilcoxon test indicated moderately significant effects between visualizations for the overall score (n = 56, U = 560, $\rho < 0.01$), situational understanding (U = 561, $\rho < 0.01$), and supply of attentional resources (U = 561, $\rho < 0.01$).

Table 4.7: SART descriptive statistics (1-low, 7-high).

	Overall and Dimensions	Mean (SD)	Median (Min/Max)
IA	Overall	4.64 (2.6)	4.5 (-1/10)
	Situational Understanding	4.96 (1.53)	5 (2/7)
	Demands on Attentional Resources	5.04 (1.2)	5 (2/7)
	Supply of Attentional Resources	4.71 (1.36)	5 (1/7)
Collective	Overall	6.68 (2.26)	6.5 (3/13)
	Situational Understanding	6.07 (0.9)	6 (4/7)
	Demands on Attentional Resources	5.07 (1.18)	5 (1/6)
	Supply of Attentional Resources	5.68 (1.09)	6 (3/7)

A summary of R_1's results that show the hypotheses with associated significant results is shown in Table 4.8. This summary table is intended to facilitate the discussion.

4.1.1.2 Discussion

Relationships to the transparency factors provided in Table 4.1 are emphasized using italics. The analysis of how visualization influenced operators suggests that the Collective visualization promoted better transparency. The variables that directly supported H_1 are the *SA performance* (i.e., accuracy) and SART. H_1 was supported, because operators using the Collective visualization had significantly higher objective and subjective *SA* and lower overall *workload*. Transparency embedded into the Collective visualization, via *explainable* color-coded icons and outlines, state information identified on the collective icon, information provided in the collective and target information pop-up windows, and feedback provided in the Collective Assignments and System Messages areas, promoted better *observability*, comprehension, and *predictability* of future collective behaviors, making the overall human-collective team more *effective*. The Collective operators; however, had more local and global clutter, even if collective pop-up win-

Table 4.8: A synopsis of R_1's hypotheses associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question. Hypotheses that were fully supported (black bold text), partially supported (black text), and not supported (gray bold text) are identified. Results for which no statistical tests were conducted are shown as merged cells containing hashmarks.

Variable	**Sub-Variable**	**Between Visualization**	**Correlation**	
			IA	Coll.
SA Probe Accuracy	SA_O	$\boldsymbol{H_1}$	———	
	SA_1	$\boldsymbol{H_1}$		
	SA_2	$\boldsymbol{H_1}$		
	SA_3	$\boldsymbol{H_1}$		
Global Clutter Percentage	SA_O	$\boldsymbol{H_1 - AT}$		
	SA_1	$\boldsymbol{H_1 - AT}$		$H_1 - B$
	SA_2	$H_1 - B, W$		
Mental Rotation Assessment	SA_O	———	$\boldsymbol{H_2}$	
	SA_1		$\boldsymbol{H_2}$	
Weekly Hours on Desktop or Laptop	SA_O		$\boldsymbol{H_2}$	
	SA_1		$\boldsymbol{H_2}$	
	SA_2			$\boldsymbol{H_2}$
NASA-TLX	Mental	$\boldsymbol{H_1}$	———	
	Performance	$\boldsymbol{H_1}$		
SART	Overall	$\boldsymbol{H_1}$		
	Situational Understanding	$\boldsymbol{H_1}$		
	Supply of Attentional Resources	$\boldsymbol{H_1}$		

dows were not considered in the local clutter calculation for the Collective visualization. Clutter was mainly attributed to the number of collective and target information pop-up windows that were visible. The increased clutter has both positive and negative implications for transparency. Clutter, from an *usability* perspective, is not ideal if operators are unable to *perform* their tasks *effectively*. The Collective operators who had more global clutter were able to answer higher *SA performance*, which suggests that operators were not hindered by the clutter and *performed* better than their counterparts. The dependence on having the collective and target information pop-up windows visi-

ble suggests that the collective state information provided on the collective icon was not as *effective* as the information pop-up window and there is a need to provide support information on the target icons. Other design strategies must be investigated to improve the efficacy of the collective and target icons for the Collective visualization. Further analyses are required to determine what contributed to more mental demand, more effort, and less perceived *performance* using the Collective visualization and whether the additional stress may have been experienced due to positive aspects, such as operators being highly motivated to complete their tasks. The Collective visualization may have required more operator effort in order to *understand* what the collective was doing compared to the IA visualization that showed the dynamic behavior emerging. Collective operators may have been distracted by the secondary *SA* probe question task and required more *time* to refocus their attention on the collective behaviors.

The Mental Rotation Assessment and video-game proficiency results supported H_2, since operators with different individual *capabilities* did not perform significantly different using the Collective visualization. One exception to the hypotheses was that more experienced operators may have *performed* better because of their extensive use of computers, which may have led to faster and more accurate interpretations of information [150] (i.e., different types of iconography), or easier access to the supplemental information. Since the exception was observed in both evaluations, the behavior is inherent to working with a computer interface, rather than a particular visualization. Using an abstract collective visualization will mitigate the need for particular operator *capabilities* to perform the sequential best-of-n decision-making task. Collective operators experienced less frustration, which supports H_3. Dis*satisfaction* (i.e., frustration) transpires when the system is not transparent and prohibits the operator from *understanding* what is happening, or there is too much clutter and the visualization appears noisy [163]. The

abstract visualization may be a solution to mitigate dis*satisfaction*.

The transparency embedded in the Collective visualization supported operators with different *capabilities* better than the IA visualization. A transparent human-collective system design will mitigate the need for operators to have particular *capabilities* in order to *effectively* interaction with the system and perform tasks successfully. Further investigation is needed to determine what visualization *usability* characteristics contributed to higher mental demand in order to alleviate *workload*.

4.1.2 R_2: Visualization Promotion of Human Operator Comprehension

The explainability direct transparency factor was encompassed in R_2, which was interested in determining whether *the visualization promoted operator comprehension*, by embedding transparency into the system design.

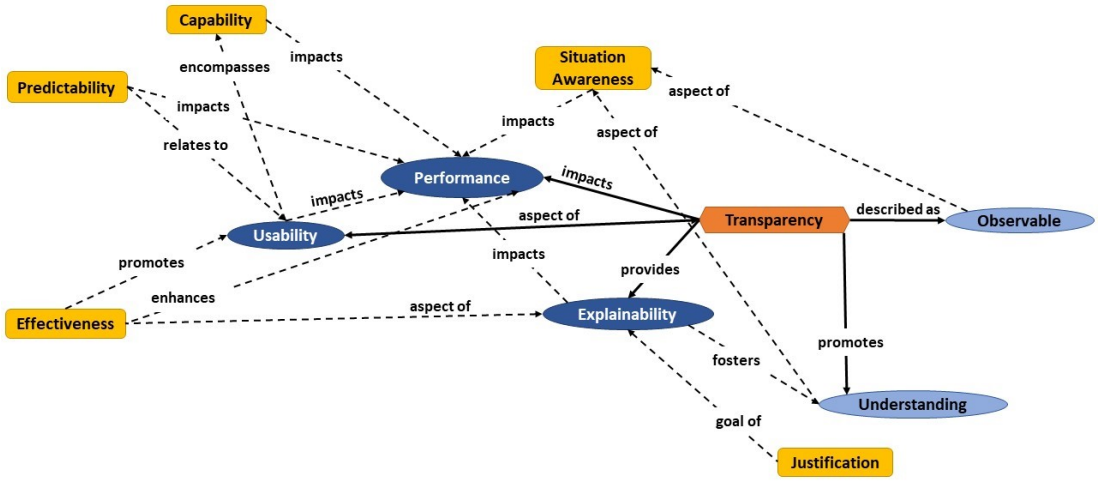


Figure 4.3: R_2 concept map of the assessed direct and indirect transparency factors.

Perception and comprehension of the presented information are necessary to inform operator actions. The associated objective dependent variables were (1) SA, (2) collective and target left- or right-clicks, (3) the percentage of times the highest value target was abandoned, and (4) whether the information pop-up window was open when a target was abandoned. The specific direct and indirect transparency factors related to R_2 are shown in Figure 4.3. The relationship between the variables and the corresponding hypotheses, and the direct and indirect transparency factors, are shown in Table 4.9.

Table 4.9: Visualization promotion of human operator comprehension objective (obj) and subjective (subj) variables (vars), relationship to the hypotheses (H), as well as the associated direct and indirect transparency factors, are presented in Figure 2.2.

		Transparency Factors									
		Direct					**Indirect**				
Obj Vars	**H**	**Explainability**	**Observable**	**Performance**	**Understanding**	**Usability**	**Capability**	**Effectiveness**	**Justification**	**Predictability**	**SA**
SA Probe Accuracy	H_4		✓	✓	✓			✓		✓	✓
Collective Left-Clicks by SA Level	H_5	✓				✓		✓	✓		
Target Right-Clicks by SA Level	H_5	✓				✓		✓	✓		
Highest Value Target Abandoned	H_4, H_5	✓			✓	✓		✓	✓		
Abandoned Target Info. Window Open	H_5	✓				✓		✓	✓		
Subj Vars											
SART	H_4				✓		✓	✓			✓
Post-Trial Performance and Understanding	H_4	✓		✓	✓		✓				

Thirteen human factors display design principles, associated with perceptual operations, mental models, human attention, and memory [150], suggest that information must be legible, clear, concise, organized, easily accessible, and consistent. Providing information, such as the collective state, on the collective icon, rather than using all of the individual collective entities is more clear, concise, organized, and consistent; therefore, it was hypothesized (H_4) that operators will have a better understanding of the information provided by the Collective visualization. Providing information redundantly via icons, colors, messages, and the collective and target information pop-up windows can aid operator comprehension and justify their future actions. It was hypothesized (H_5) that the Collective visualization provided information used to accurately justify actions. An ideal visualization will enable operators to perceive and comprehend information that is explainable, which will support taking accurate future actions.

4.1.2.1 Metrics and Results

The operator had access to supplementary information that was not continually displayed, such as different colored target borders that identified which targets were in range and had been abandoned, or information pop-up windows that provided collective state and target support information, in order to aid comprehension (SA_2) of collective behavior and inform particular actions. The results of *SA probe accuracy*, which is the percentage of correctly answered SA probes questions used to assess the operator's SA during a trial, identified that operators using the Collective visualization had higher SA probe accuracy, regardless of the SA level. Further details regarding the statistical tests were provided in Chapter 4.1.1.1.

Collective left-clicks identified targets that were in range of a collective (i.e., white borders indicated that the individual collective entities were investigating the target, while yellow indicated no investigation), whether the targets had been abandoned (i.e., red borders), and was the first click required to issue a command. The number of collective left-clicks descriptive statistics 15 seconds before asking, while being asked, and during response to a SA probe question are shown in Table 4.10. Operators using the

Table 4.10: Collective left-clicks descriptive statistics 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	**SA Level**	**Mean (SD)**	**Median (Min/Max)**
IA	Before	SA_O	1.64 (1.84)	1 (0/12)
		SA_1	1.53 (1.75)	1 (0/11)
		SA_2	1.78 (1.9)	1 (0/12)
		SA_3	1.65 (1.92)	1 (0/9)
	Asking	SA_O	0.49 (0.76)	0 (0/5)
		SA_1	0.3 (0.6)	0 (0/3)
		SA_2	0.42 (0.77)	0 (0/4)
		SA_3	0.33 (0.61)	0 (0/3)
	Responding	SA_O	1.68 (1.79)	1 (0/11)
		SA_1	1.14 (1.46)	1 (0/7)
		SA_2	1.46 (1.8)	1 (0/10)
		SA_3	1.53 (1.98)	1 (0/9)
Collective	Before	SA_O	1.95 (1.57)	2 (0/9)
		SA_1	1.88 (1.47)	2 (0/8)
		SA_2	2.13 (1.68)	2 (0/9)
		SA_3	1.83 (1.61)	1 (0/7)
	Asking	SA_O	0.69 (0.88)	0 (0/5)
		SA_1	0.51 (0.79)	0 (0/5)
		SA_2	0.91 (0.89)	1 (0/4)
		SA_3	0.73 (0.96)	0 (0/4)
	Responding	SA_O	1.52 (1.21)	1 (0/6)
		SA_1	1.32 (1.02)	1 (0/4)
		SA_2	1.57 (1.21)	1 (0/5)
		SA_3	1.89 (1.48)	2 (0/6)

IA visualization had fewer collective left-clicks, regardless of when the metric was assessed for all SA levels, except for SA_O during response to a SA probe question. The Mann-Whitney-Wilcoxon tests found highly significant effects between visualizations for SA_O (n = 664) 15 seconds before asking (U = 64213, $\rho < 0.001$), while being asked (U = 67670, $\rho < 0.001$), and during response to a SA probe question (U = 64710, $\rho < 0.001$). A highly significant effect was also found when responding to a SA probe question for SA_2 (n = 223, U = 8317, $\rho < 0.001$). Moderate significant effects were found for SA_1 (n = 290) 15 seconds before asking (U = 12534, $\rho < 0.01$), while being asked (U = 12043, $\rho < 0.01$), and during response to a SA probe question (U = 12414, $\rho < 0.01$). An additional moderate significant effect was found while being asked a SA probe question for SA_3 (n = 151, U = 3472, $\rho < 0.01$). Significant effects were found 15 seconds before asking a SA probe question for SA_2 (U = 7210.5, $\rho = 0.04$) and during response to a SA probe question for SA_3 (U = 3489, $\rho = 0.01$). No correlations were found between the number of collective left-clicks and SA probe accuracy.

Target right-clicks allowed the operator to open or close target information pop-up windows, which provided the percentage of support each collective had for a respective target. Operators may have used the support information to justify issuing commands. The number of target right-clicks descriptive statistics 15 seconds before asking, while being asked, and during response to a SA probe question are presented in Table 4.11. The Collective visualization had fewer target right-clicks for all SA levels, 15 seconds before asking and during response to a SA probe question, and the IA visualization had fewer while being asked a SA probe question. The Mann-Whitney-Wilcoxon test found no significant effects between visualizations for the number of target right-clicks. The Spearman correlation analysis revealed weak correlations between the number of target right-clicks and SA probe accuracy for the IA visualization 15 seconds before asking a

SA probe question for SA_O (r = 0.17, $\rho < 0.01$) and SA_2 (r = 0.37, $\rho < 0.001$).

Table 4.11: Target right-clicks descriptive statistics 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	**SA Level**	**Mean (SD)**	**Median (Min/Max)**
IA	Before	SA_O	1.68 (2.38)	1 (0/13)
		SA_1	1.92 (2.62)	1 (0/13)
		SA_2	1.28 (1.91)	1 (0/11)
		SA_3	1.8 (2.5)	1 (0/12)
	Asking	SA_O	0.37 (0.79)	0 (0/7)
		SA_1	0.44 (0.74)	0 (0/4)
		SA_2	0.31 (0.67)	0 (0/3)
		SA_3	0.37 (0.75)	0 (0/3)
	Responding	SA_O	1.07 (1.77)	0 (0/10)
		SA_1	1.11 (1.69)	0 (0/10)
		SA_2	1.1 (1.75)	0 (0/9)
		SA_3	1.68 (2.24)	1 (0/10)
Collective	Before	SA_O	1.52 (2.41)	1 (0/18)
		SA_1	1.79 (2.71)	1 (0/18)
		SA_2	1.17 (1.94)	0 (0/10)
		SA_3	1.49 (2.35)	0 (0/11)
	Asking	SA_O	0.5 (1)	0 (0/9)
		SA_1	0.49 (0.86)	0 (0/4)
		SA_2	0.55 (1.31)	0 (0/9)
		SA_3	0.44 (0.69)	0 (0/2)
	Responding	SA_O	0.99 (1.7)	0 (0/11)
		SA_1	1.01 (1.74)	0 (0/11)
		SA_2	0.84 (1.44)	0 (0/9)
		SA_3	1.21 (1.98)	0 (0/8)

The abandon command was provided to operators who desired a collective to discontinue investigating a particular target. Ideally lower valued targets were abandoned, since the objective was to aid each collective in selecting and moving to the highest valued target two sequential times. The percentage of times the *highest value target was abandoned* per participant is presented in Table 4.12. Operators using the IA

visualization abandoned the highest value target less frequently, but no significant effects were found between the visualizations.

Table 4.12: Highest value target abandoned (%) descriptive statistics per participant by decision difficulty.

	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	Overall	32.36 (29.53)	26 (0/100)
	Easy	31.2 (27.17)	25 (0/75)
	Hard	42.1 (40.53)	23 (0/100)
Collective	Overall	43.6 (31.94)	38 (0/100)
	Easy	33.25 (35.96)	27 (0/100)
	Hard	48.72 (36.85)	38 (0/100)

The instances when a target was abandoned by an operator and the target's respective information pop-up window was visible was assessed. The *abandoned target information pop-up window was open* per participant is presented in Table 4.13. The operator may have used the support information in order to justify abandoning a target. IA operators had fewer abandoned target information pop-up windows open, compared to the Collective operators. No significant effects were found between visualizations.

Table 4.13: Abandoned target information pop-up window open (%) descriptive statistics per participant by decision difficulty.

	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	Overall	23.86 (31.43)	10.5 (0/100)
	Easy	22.2 (30.95)	13 (0/100)
	Hard	28.7 (37.89)	8.5 (0/100)
Collective	Overall	33.8 (34.9)	15 (0/100)
	Easy	30.7 (37.85)	15 (0/100)
	Hard	36.08 (40.87)	14 (0/100)

The *SART* results, which measured the operator's perceived situational understanding, demand on attentional resources, and supply of attentional resources [175], were ranked higher for the Collective visualization compared to the IA. The statistical test

details were provided in Chapter 4.1.1.1.

The post-trial questionnaire assessed the operators' *understanding of the collective behavior*, never (1) to always (7), and their *ability* to chose the best target for each decision, never (1) to always (7). The post-trial performance and understanding subjective ranking descriptive statistics are presented in Table 4.14 [29]. The performance and understanding rankings were higher for operators using the Collective visualization. The Mann-Whitney-Wilcoxon test found a significant effect between visualizations for understanding (n = 56, U = 513, ρ = 0.04).

Table 4.14: Post-trial performance and understanding model ranking descriptive statistics (1-low, 7-high).

	Metric	**Mean (SD)**	**Median (Min/Max)**
IA	Performance	5.25 (1.69)	6 (2/7)
	Understanding	4.89 (1.75)	5 (1/7)
Collective	Performance	5.54 (1.29)	6 (3/7)
	Understanding	5.82 (1.16)	6 (3/7)

A summary of R_2's results that show the hypotheses with associated significant results is shown in Table 4.15. This summary table is intended to facilitate the discussion.

4.1.2.2 Discussion

The analysis of how visualization promoted operator comprehension (i.e., the operator's *capability* of *understanding*) identified advantages and disadvantages associated with both visualizations. The Collective visualization promoted higher comprehension and *SA*; however, because the Collective operators abandoned the highest value target more frequently, H_4 was not supported. In*effective* identifiers, such as the distinction between roman numerals and integers, may have caused poor *observability* and IA op-

Table 4.15: A synopsis of R_2's hypotheses associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question.

Variable	**Sub-Variable**	**Between Visualization**	**Correlation**	
			IA	Coll.
SA Probe Accuracy	SA_O	$\boldsymbol{H_4}$	———	
	SA_1	$\boldsymbol{H_4}$		
	SA_2	$\boldsymbol{H_4}$		
	SA_3	$\boldsymbol{H_4}$		
Collective Left-Clicks	SA_O	$H_5 - AT$		
	SA_1	$H_5 - AT$		
	SA_2	$H_5 - B, W$		
	SA_3	$H_5 - W, D$		
Target Right-Clicks	SA_O		$H_5 - B$	
	SA_2		$H_5 - B$	
SART	Overall	$\boldsymbol{H_4}$	———	
	Situational Understanding	$\boldsymbol{H_4}$		
	Supply of Attentional Resources	$\boldsymbol{H_4}$		
Post-Trial	Understanding	$\boldsymbol{H_4}$		

erator confusion. Ensuring that identifiers are unique and distinct, such as integers versus letters, will improve system *explainability* and may mitigate mis*understanding*. The target value for the Collective visualization may not have been salient enough to distinguish it from other potential targets. Further investigations are required to determine if the target value must use the entire collective hub icon area, similar to the IA visualization, in order to be recognizable, and to establish what levels of obscurity are needed to ensure that target values are distinguishable from one another.

The use of target borders (collective left-clicks), information pop-up windows (target right-clicks), and target value, were assessed to determine if operators used these types of information to accurately *justify* actions. Collective right-clicks, which opened and closed collective information pop-up windows, 15 seconds before asking, while

being asked, and during response to a *SA* probe question, were not analyzed in the Visualization Analysis, because the collective evaluation did not indicate which collective was selected via the right-click. None of the metrics supported H_5, because the information provided by the Collective visualization did not *justify* accurate actions. Collective left-clicks did not support or hinder *SA performance* (i.e., accuracy) for either visualization, but fewer target right-clicks 15 seconds before asking a *SA* probe question supported higher SA_O and SA_2 probe accuracy for operators using the IA visualization. The operators may have learned to anticipate when *SA* probe questions were going to be asked and took preventative actions, by opening or closing target information windows, which resulted in higher *SA performance*. The use of target information pop-up windows aided Collective visualization users to abandon targets more than 30% of the time. Further analysis using technology, such as an eye-tracker, may provide more accurate metrics to determine operator comprehension during *SA* probe questions by identifying exactly where an operator is focusing their attention.

The transparency embedded in the Collective visualization did not provide the operator with better comprehension, nor did it hinder it compared to the IA visualization. Both visualizations influenced operator comprehension differently. Further investigations are needed in order to *understand* how to embed transparency into the system better and identify whether those strategies worked. Unique and distinct information may need to be presented on different icons, or use different presentation strategies, such as colors versus patterns, in order to mitigate confusion and improve system *usability*. The target value on the Collective visualization, for example, was on the same icon as the collective support, and may have challenged operator perception, comprehension, and *predictability*. The information provided by both of the visualizations in general did not *justify* actions.

4.1.3 R_3: Visualization Usability

Understanding *which visualization promoted better usability*, R_3, is necessary to determine which system characteristics promote effective transparency in human-collective systems. The associated objective dependent variables were (1) visualization clutter, (2) Euclidean distance, (3) collective and target left- and right-clicks, (4) metrics associated with abandoned targets, and (5) the time between the commit state and issued decide command. The specific direct and indirect transparency factors related to R_3 are shown in Figure 4.4. The relationship between the variables and the corresponding hypotheses, as well as the direct and indirect transparency factors are shown in Table 4.16.

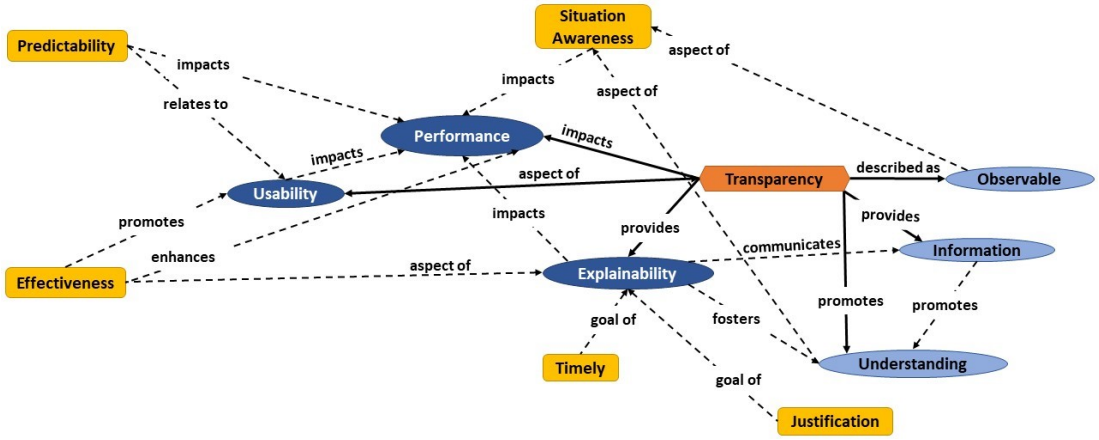


Figure 4.4: R_3 concept map of the assessed direct and indirect transparency factors.

The goal of usability is to design systems that are effective, efficient, safe, have good utility, easy to learn, and are memorable [163]. Ensuring good usability is necessary to ensure operators will be able to perceive and understand the information presented on a visualization, and to promote effective interactions. It was hypothesized (H_6) that the Collective visualization will promote better usability by being more predictable and ex-

Table 4.16: Visualization usability objective (obj) variables (vars), relationship to the hypotheses (H), as well as the associated direct and indirect transparency factors, are presented in Figure 2.2.

		Transparency Factors										
		Direct						**Indirect**				
Obj Vars	**H**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Effectiveness**	**Justification**	**Predictability**	**SA**	**Timing**
Local Clutter	H_6			✓	✓	✓	✓	✓		✓	✓	
Global Clutter	H_6			✓	✓	✓	✓	✓		✓	✓	
Euclidean Distance Between SA Probe	H_7						✓	✓				
Interest and Clicks												
Sum of Euclidean Distance Between Clicks	H_7						✓	✓				
Collective Left-Clicks per Participant	H_7						✓	✓	✓			
Collective Right-Clicks per Participant	H_7		✓				✓	✓	✓			
Target Left-Clicks per Participant	H_7						✓	✓				
Target Right-Clicks per Participant	H_7		✓				✓	✓	✓			
Highest Value Target Abandoned	H_6	✓				✓	✓	✓	✓			
Abandoned Target Info. Window Open	H_6	✓	✓				✓	✓	✓			
Abandon Requests Exceeded Abandoned Targets	H_6	✓				✓	✓	✓				
Time Between Commit State and Issued Decide Command	H_6			✓			✓	✓		✓	✓	✓

plainable. Providing information that is explainable may aid operator comprehension, while predictable information may expedite operator actions. An ideal system will not require constant operator interaction to perform well; therefore, it was hypothesized (H_7) that operators using the Collective visualization will require fewer interactions.

4.1.3.1 Metrics and Results

Many system characteristics were available to the operators in order to aid task completion. The IA visualization had both lower *local clutter percentage*, which was the percentage of area obstructed by items displayed within the 500 *m* circular radius of a collective, or target, and *global clutter percentage*, which was the percentage of area obstructed by all objects displayed on the visualization. Operators using the IA visualization had fewer collective and target information pop-up windows open throughout the trial. The statistical test details were provided in Chapter 4.1.1.1.

The Euclidean *distance (pixels) between the focus of the SA probe question and where the operator was interacting* with the visualization indicated where operators focused their attention, because no eye-tracker was used. Euclidean distance can be used to assess the effectiveness of the object placements on the display. Larger distances are not ideal, because more time [176] and effort is required to locate and interact with the object. The first requirement of calculating the Euclidean distance was to determine what the collective, or target of interest was in a SA probe question. For example, Target 3 is the target of interest for the following question: "What collectives are investigating Target 3?" The second requirement was to determine where the operator was interacting with the system. The Euclidean distance between Target 3 and the operator's current interaction (e.g., click), which was Collective IV, is identified by a dashed orange line in Figure 4.5.

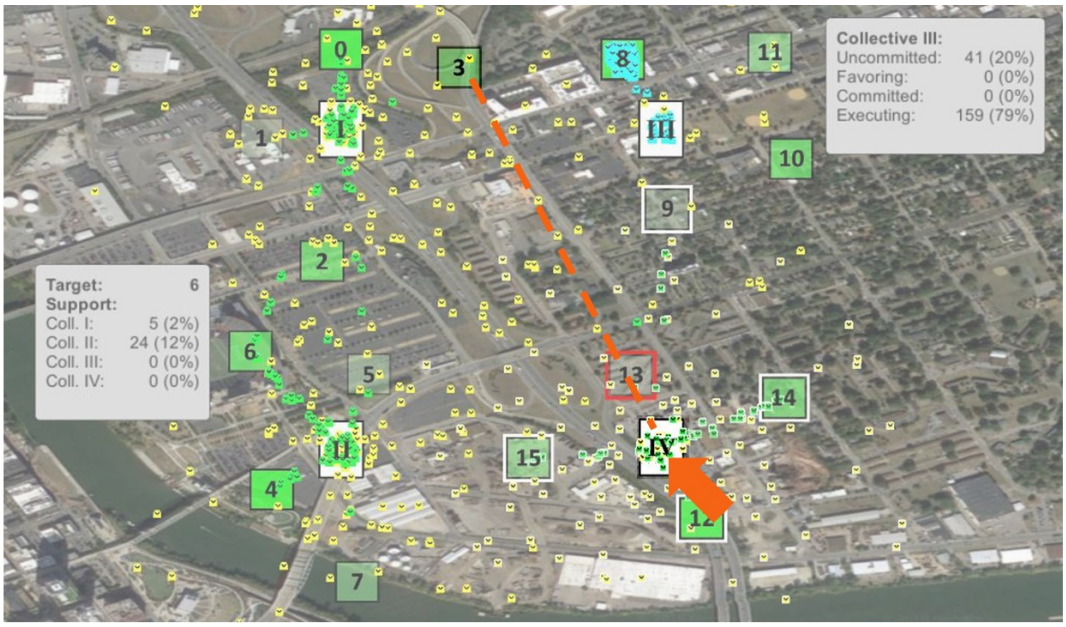


Figure 4.5: Example of Euclidean distance between SA probe interest (Target 3) and clicks (Collective IV), denoted by an orange dashed line.

The Euclidean distance between SA probe interest and clicks descriptive statistics 15 seconds before asking, while being asked, and during response to a SA probe question are presented in Table 4.17. Operators using the IA visualization had shorter Euclidean distances between the SA probe interest and their visualization-based clicks, regardless of when the metric was assessed for all SA levels. The Mann-Whitney-Wilcoxon tests found a moderate significant effect between visualizations while being asked a SA probe question for SA_O (n = 464, U = 31052, $\rho < 0.01$). Highly significant effects were found between visualizations 15 seconds before asking a SA probe question for SA_O (n = 557, U = 43303, $\rho = 0.02$) and SA_1 (n = 273, U = 10577, $\rho = 0.05$), while being asked a SA probe question for SA_1 (n = 229, U = 7645, $\rho = 0.01$), and during response to a SA probe for SA_O (n = 499, U = 35029, $\rho = 0.02$). The Spearman correlation analysis revealed a

weak correlation between the Euclidean distance of the SA probe's focus interest to the current clicks and SA probe accuracy for the IA visualization 15 seconds before asking a SA probe question for SA_1 (r = -0.18, ρ = 0.04).

Table 4.17: Euclidean distance between SA probe interest and clicks (pixels) descriptive statistics 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	**SA Level**	**Mean (SD)**	**Median (Min/Max)**
IA	Before	SA_O	767.1 (262.5)	768.3 (183.4/1425.9)
		SA_1	759.5 (251.64)	765.1 (269/1425.9)
		SA_2	768.9 (282.07)	787.9 (184.4/1312.4)
		SA_3	783.4 (262.89)	757.4 (183.4/1260.5)
	Asking	SA_O	758.44 (291.48)	744.91 (73.72/1636.46)
		SA_1	754.4 (284.65)	744.6 (139.7/1636.5)
		SA_2	768.4 (316.09)	782.4 (160.2/1382.8)
		SA_3	753.7 (275.04)	730.68 (73.72/1329.14)
	Responding	SA_O	764.24 (298.84)	757.53 (73.72/1636.46)
		SA_1	760.9 (297.14)	746.6 (160.6/1636.5)
		SA_2	774.6 (319.08)	777.6 (160.2/1381.2)
		SA_3	757.71 (278.14)	749.85 (73.72/1283.9)
Collective	Before	SA_O	820.7 (255.67)	855.9 (222.3/1470.5)
		SA_1	825.6 (264.1)	828.1 (249.8/1470.5)
		SA_2	812.9 (234.94)	865.2 (317.4/1329.9)
		SA_3	821.6 (271.03)	873.1 (222.3/1243.9)
	Asking	SA_O	851.4 (293.91)	859.5 (280.2/1745.2)
		SA_1	845.5 (282.53)	862.5 (281.5/1745.2)
		SA_2	879.5 (299.93)	869.2 (280.2/1696.1)
		SA_3	823.5 (314.47)	787.6 (314.7/1469)
	Responding	SA_O	827.7 (273.83)	819.8 (258.1/1546)
		SA_1	827.9 (279.21)	819 (366.6/1484.6)
		SA_2	845.2 (275.55)	862.7 (258.1/1546)
		SA_3	799.7 (261.1)	797.3 (314.7/1378.8)

The *sum Euclidean distance (pixels) between clicks* was the sum of all distances between the operator's current interaction and the immediately previous interaction. For example, if an operator interacted with the visualization four times while a SA probe

question was being asked, the sum Euclidean distance is the sum between interactions one and two, interactions two and three, and interactions three and four. The sum of Euclidean distance between clicks descriptive statistics 15 seconds before asking, while being asked, and during response to a SA probe question are presented in Table 4.18. Operators using the Collective visualization had smaller sums of Euclidean distance between their interactions, regardless of when the metric was assessed for all SA lev-

Table 4.18: Sum of Euclidean distance between clicks (pixels) descriptive statistics 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	**SA Level**	**Mean (SD)**	**Median (Min/Max)**
IA	Before	SA_O	3386 (2911.78)	2500 (0/21115)
		SA_1	3185 (2665.35)	2476 (0/18465)
		SA_2	3389 (3091.22)	2469 (0/21115)
		SA_3	3732.6 (3071.18)	2901.7 (246.1/17323.6)
	Asking	SA_O	1748.9 (1779.6)	1250.7 (0/12161.7)
		SA_1	1387.3 (1328.49)	985.7 (0/6777.4)
		SA_2	2030.1 (2053.98)	1477.5 (0/12161.7)
		SA_3	1975 (1954.76)	1362 (0/8949)
	Responding	SA_O	1383.2 (1416.84)	934.8 (0/9315.1)
		SA_1	1088.7 (1108.12)	793.5 (0/4999)
		SA_2	1606 (1610.13)	1310 (0/9315)
		SA_3	1582.6 (1534.29)	1211.9 (0/6728.2)
Collective	Before	SA_O	3187 (1983.72)	2834 (0/9690)
		SA_1	3085 (1911.3)	2611 (0/8898)
		SA_2	3235 (2049.38)	3041 (0/9690)
		SA_3	3331.6 (2047.61)	3097.7 (195.8/8910.7)
	Asking	SA_O	1741.3 (1323.11)	1434.6 (0/6323.2)
		SA_1	1778 (1437)	1380 (0/5920)
		SA_2	1693 (1093.21)	1619 (0/4411)
		SA_3	1743.8 (1439.54)	1371 (0/6323.2)
	Responding	SA_O	1218.5 (944.41)	1047.1 (0/4428.2)
		SA_1	1174.7 (992.56)	995 (0/4277.6)
		SA_2	1176.8 (801.24)	1074.6 (0/4084.9)
		SA_3	1373.15 (1041)	1160.81 (65.35/4428.24)

els, with two exceptions. The IA visualization had a smaller sum for SA_1 while being asked and during response to a SA probe question. The Mann-Whitney-Wilcoxon test found no significant effects between visualizations. The Spearman correlation analysis revealed weak correlations between the sum of Euclidean distance between clicks and SA probe accuracy for the IA visualization 15 seconds before asking a SA probe question for SA_2 (r = 0.2, ρ = 0.04) and during response to a SA probe question for SA_O (r = 0.14, ρ = 0.02). Weak correlations were revealed for the Collective visualization while being asked a SA probe question for SA_O (r = -0.13, ρ = 0.05) and SA_1 (r = -0.2, ρ = 0.05).

Collective and target left- and right-clicks were examined per participant. *Target left-clicks* were the second click required in the process of issuing commands, but did not provide supplementary information. The number of collective and target left- and right-clicks descriptive statistics are presented in Table 4.19. Operators using the IA visualization had fewer collective and target left-clicks, while those using the Collective visualization had fewer collective and target right-clicks. The Mann-Whitney-Wilcoxon test identified no significant effects between visualizations.

Table 4.19: Collective and target left- and right-clicks per participant descriptive statistics.

	Clicks	**Mean (SD)**	**Median (Min/Max)**
IA	Collective Left	107.6 (49.89)	104 (5/235)
	Collective Right	30.64 (20.98)	27.5 (0/85)
	Target Left	97.64 (58.78)	83 (5/251)
	Target Right	97.18 (82.79)	68.5 (4/352)
Collective	Collective Left	121.96 (47.4)	130.5 (35/212)
	Collective Right	30.57 (31.95)	19.5 (7/164)
	Target Left	185.6 (64.32)	202 (62/290)
	Target Right	82.39 (60.22)	75 (23/278)

Metrics showing how operators used the abandon command were assessed. IA operators had lower percentages of times the *highest value target was abandoned* and lower percentages of times an *abandoned target information pop-up window was open* per participant. The statistical analyses of both metrics were provided in Chapter 4.1.2.1. Instances may have occurred when the operator accidentally issued an undesired abandon command or repeatedly issued the abandon command, although the command only needed to be issued once; hence, the percent of times *abandon commands exceeded abandoned targets* was examined and the descriptive statistics are shown in Table 4.20. Operators using the IA visualization had fewer repeated abandon commands for all decision difficulties, compared to those using the Collective visualization. The Mann-Whitney-Wilcoxon test found no significant effects between visualizations.

Table 4.20: The percentage of times abandon commands exceeded abandoned targets per participant descriptive statistics.

	Decision Difficulty	**Mean (SD)**	**Median (Min/Max)**
IA	Overall	1.18 (3.02)	0 (0/12)
	Easy	0.4 (1.55)	0 (0/6)
	Hard	1.35 (4)	0 (0/16)
Collective	Overall	2.68 (6.27)	0 (0/22)
	Easy	2.05 (5.06)	0 (0/15)
	Hard	3.08 (7.74)	0 (0/25)

A collective's entities moved to the operator selected target when the decide command was issued. A decide request required at least 30% of the collective support for the operator specified target. Collectives that reached 50% support for a target transitioned into the executing state and the operator was no longer able to influence the collective behavior. The *time difference (minutes) between the committed state and issued decide command* assessed the operator's ability to predict the collective's future state changing from the committed state (30% support for a target) to executing (50% support for a

target). The time difference between the committed state and when an operator issued a decide command descriptive statistics are shown in Table 4.21. Operators using the Collective visualization had smaller time differences between the committed state and issued decide commands for overall and easy decisions; however, operators using the IA visualization had smaller time differences for hard decisions. The Mann-Whitney-Wilcoxon test found no significant effects between visualizations.

Table 4.21: The time difference (minutes) between committed state and issued decide request per participant descriptive statistics.

	Decision Difficulty	**Mean (SD)**	**Median (Min/Max)**
IA	Overall	0.68 (0.27)	0.62 (0.42/1.78)
	Easy	0.7 (0.47)	0.63 (0.32/2.56)
	Hard	0.72 (0.21)	0.66 (0.41/1.15)
Collective	Overall	0.65 (0.15)	0.63 (0.45/1.18)
	Easy	0.56 (0.14)	0.58 (0.27/0.88)
	Hard	0.78 (0.3)	0.75 (0.47/1.99)

A summary of R_3's results that show the hypotheses with associated significant results is shown in Table 4.22. This summary table is intended to facilitate the discussion.

Table 4.22: A synopsis of R_3's hypotheses associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question.

Variable	**Sub-Variable**	**Between Visualization**	**Correlation**	
			IA	Coll.
Global Clutter Percentage	SA_O	$H_6 - AT$		
	SA_1	$H_6 - AT$		$H_6 - B$
	SA_2	$H_6 - B, W$		
Euclidean Distance Between SA Probe Interests and Clicks	SA_O	$H_7 - AT$		
	SA_1	$H_7 - B, W$	$H_7 - B$	
Sum of Euclidean Distance Between Clicks	SA_O		$H_7 - D$	$H_7 - W$
	SA_1			$H_7 - W$
	SA_2		$H_7 - B$	

4.1.3.2 Discussion

The analysis of which visualization promoted better *usability* was inconclusive, because both visualizations had advantages and disadvantages. The Collective visualization's *predictability justified* operators issuing decide commands faster, after the collective was in the committed state, compared to those using the IA visualization. H_6 was not supported; however, because the Collective operators abandoned the highest value target more frequently, and there was a higher percentage of abandon commands, which exceeded the number of abandoned targets. Collective operators had more local and global clutter, which suggests that Collective operators relied on the *information* pop-up windows to answer the *SA* probe questions more than IA operators. Sixteen of twenty-four *SA* probe questions relied on *information* provided in the *information* pop-up windows. The collective and target icons, as well as the target outlines, were intended to aid Collective operators to answer the SA probe questions correctly; however, the operators needed to use the *information* pop-up windows in order to see the numeric values for the collective support and behaviors in order to answer questions regarding target support from a specific collective, or multiple collectives. An example question, such as "What collectives are investigating Target 3?" in Figure 3.2, will require using the target *information* pop-up window, because Target 3 is in range of Collective I and III. A target *information* pop-up window is not required for Target 1, in Figure 3.2, since it is only in range of Collective I. The need to use *information* pop-up windows contributed to the Collective visualization clutter. The operators may have preferred the numeric value representations (e.g., more *explainable*) versus the other visualization techniques, which may have contributed to their reliance on the *information* pop-up windows.

The IA operators may have had an advantage, by deducing the same *information* as the Collective operators gained from the *information* pop-up windows, by *observing* the dynamic behavior of the individual collective entities. Relying on supplemental *information* pop-up windows is not ideal and suggests that improvements must be made to the collective icon to ensure the collective's state *information* is more *understandable*. *Usability* modifications, such as indicating which collective was the highest supporting collective on the target icon, instead of only showing that there was support through the use of color and opacity, may increase the target icon's *effectiveness* and reduce reliance of the target *information* pop-up window. Additional experimental design modifications can ensure a more even distribution of questions that may rely on other *information* providing visualization features, such as the icons, system messages, or collective assignments versus *information* pop-up windows. Operators using target *information* pop-up windows to verify that a target was abandoned by a collective, may have been confused if the reported target support was greater than zero. The operators may have reissued additional abandon commands in order to reduce the collective support to zero, although only one abandon command was needed. There were instances during the trial when a few individual collective entities became lost, when the collective hub was transitioning to a new location, and never moved with the hub. The lost entities may have continued to explore a now abandoned target, because they never received the message to abandon the target, which was only communicated inside of the hub. Strategies, such as reporting zero percent support when an abandon command is issued, may help mitigate erroneous issuing of repeated abandon commands, which was experienced up to 25% for some operators. IA operators may have also experienced confusion if they saw individual collective entities still travelling to an abandoned target. Not displaying lost entities after a specific period of *time* once a collective hub has

moved to a new location may also reduce the number of reissued abandon commands.

SA_1 probe questions that inquired about objects nearby were answered more accurately than those that were further away when using the IA visualization. Asking SA probe questions about objects at various distances from the operator's current focal point is necessary in order to *understand* how clutter, or moving individual collective entities, may affect the operator's ability to identify the object of interest and answer the question correctly. Smaller sum of Euclidean distances between interactions, suggests Collective operators may have had fewer interactions. Further analysis is required to determine whether more interactions were needed for operators to answer *SA* probe questions correctly and improve decision *performance*. H_7 was not supported by the analysis. The IA operators may have issued fewer commands, or did not rely on target borders as much as the Collective operators. Issuing more commands suggests that Collective operators may have wanted more control over the decision-making task, which may have occurred due to lower trust, or mis*understanding* collective behavior. Further investigations are needed in order to *understand* how the model and control mechanisms may interact with the visualization of collectives to influence operator behavior. The *effectiveness* of the system design will be dependent on all system characteristics working together to promote optimal human-collective *performance*.

The transparency embedded in the Collective visualization did not support the best overall system *usability*. The IA visualization promoted less clutter by alleviating the need to use the collective and target *information* pop-up windows as often, and promoted fewer interactions. *Usability* and *explainability* modifications to the Collective visualization are needed in order to mitigate undesired operator behaviors, such as the highest value target being abandoned more frequently, as well as reduce the reliance on the *information* windows. Transparency embedded into the Collective visualization,

via the collective hub icons for example, must represent the same types of information provided in the information windows. The assumption that fewer interactions are optimal may not be accurate for all situations. *Understanding* strategies and *justifications* for more interactions is necessary in order to promote transparency that aids operators during particular situations and results in higher human-collective *performance*.

4.1.4 R_4: Visualization Influence on Human-Collective Performance

Assessing *which visualization promoted better human-collective performance*, R_4, is necessary to determine whether the human-collective system transparency aided task completion. An ideal system performs a task quickly, safely, and successfully.

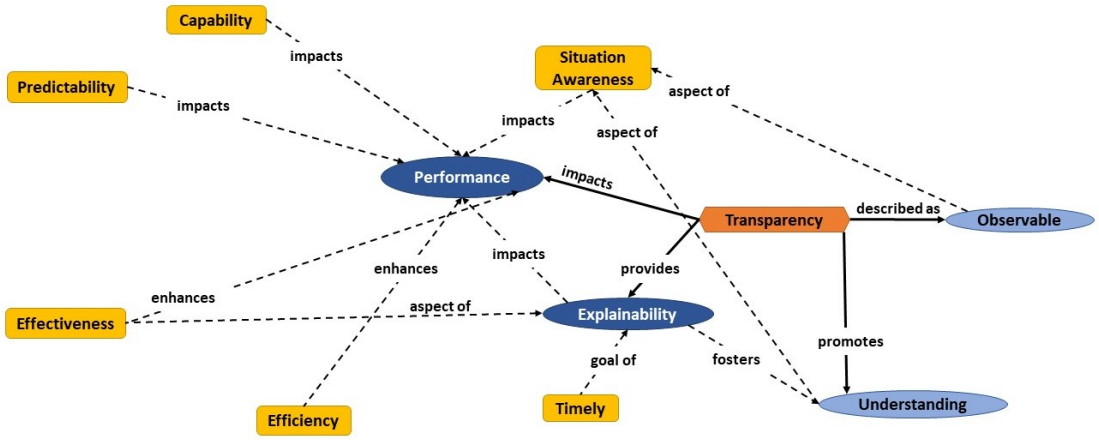


Figure 4.6: R_4 concept map of the assessed direct and indirect transparency factors.

The associated objective dependent variables were (1) decision time, (2) selection success rate, and (3) SA probe accuracy. The specific direct and indirect transparency factors related to R_4 are identified in Figure 4.6. The relationship between the variables and the corresponding hypotheses, as well as the direct and indirect transparency fac-

tors, are identified in Table 4.23.

Performance of the human-collective team can be used to assess the effects of visualization transparency on the team's ability to fulfill tasks. An ideal system promotes high performance rates. It was hypothesized (H_8) that the human-collective performance, effectiveness, efficiency, and timing will be better using the Collective visualization.

Table 4.23: Visualization influence on human-collective performance objective (obj) and subjective (subj) variables (vars), relationship to the hypothesis (H_8), as well as the associated direct and indirect transparency factors, are presented in Figure 2.2.

	Transparency Factors									
	Direct				**Indirect**					
Obj Vars	**Explainability**	**Observable**	**Performance**	**Understanding**	**Capability**	**Effectiveness**	**Efficiency**	**Predictability**	**SA**	**Timing**
Decision Time Per Decision			✓			✓	✓			✓
Selection Success Rate Per Decision			✓			✓				
SA Probe Accuracy		✓	✓	✓		✓		✓	✓	
Mental Rotation Assessment			✓		✓					
Subj Vars										
Weekly Hours on a Desktop or Laptop			✓		✓					
Video Game Proficiency			✓		✓					
Post-Trial Performance and Understanding	✓		✓	✓	✓					

4.1.4.1 Metrics and Results

The length of time it took the human-collective team to reach a decision, *decision time* (minutes), was examined. Consensus decision-making algorithms are inherently slow,

which is undesirable in realistic use scenarios. Adding a human operator into the loop permits the human to influence the decision and has the potential to minimize decision time. The decision time descriptive statistics per decision are shown in Table 4.24 [57, 171]. Operators using the Collective visualization had faster decision times for overall, easy, and hard decisions. Both visualizations had faster human-collective decision times compared to the simulation (M_{2SIM}). The Collective visualization simulation had slightly faster decision times for overall and easy decisions, while the IA visualization simulation had faster decision times for hard decisions. The Mann-Whitney-Wilcoxon test found significant effects between visualizations with human operators for overall (n = 672, U = 50921, ρ = 0.03), easy (n = 375, U = 15452, ρ = 0.04), and hard decisions (n = 297, U = 9521, ρ = 0.04). Highly significant effects were found between human operators and simulation for the IA visualization overall (U = 74005, $\rho < 0.001$), easy (n = 481, U = 37414, $\rho < 0.001$), and hard decisions (n = 384, U = 23194, $\rho < 0.001$) and for the Collective visualization overall (U = 79468, $\rho < 0.001$), easy (n = 461, U = 38786, $\rho <$ 0.001), and hard decisions (n = 392, U = 26887, $\rho < 0.001$).

Table 4.24: Decision time (minutes) descriptive statistics per decision difficulty.

	Model	**Decision Difficulty**	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	Overall	4.32 (1.83)	3.94 (1.74/15.94)
		Easy	3.77 (1.63)	3.38 (1.74/13.47)
		Hard	5.09 (1.82)	4.68 (1.86/15.94)
	M_{2SIM}	Overall	4.8 (1.1)	4.82 (2.46/7.68)
		Easy	4.19 (1.06)	4.07 (2.46/8.85)
		Hard	5.73 (1.26)	5.54 (3.43/10.15)
Collective	M_2	Overall	3.97 (1.37)	3.64 (1.83/9.94)
		Easy	3.37 (1.23)	3.09 (1.83/9.94)
		Hard	4.67 (1.2)	4.57 (2.46/8.81)
	M_{2SIM}	Overall	4.79 (1.11)	4.79 (2.49/7.7)
		Easy	4.17 (0.93)	4.1 (2.49/7.55)
		Hard	5.77 (1.38)	5.62 (3.67/10.25)

The *selection success rate* was the number of correct decisions (the collective moved to the highest valued target based on selecting the target itself or due to the operator issuing a decide command)) relative to the total number of decisions. Selection success rate descriptive statistics per decision are shown in Table 4.25 [57, 171]. Collective operators had higher selection success rates for all decision difficulties. The human-collective teams had higher selection success rates for both visualizations for all decision difficulties compared to the simulation. The Collective simulation had higher selection success rates for overall and hard decisions, while the IA simulation had higher selection success rates for easy decisions. The Mann-Whitney-Wilcoxon test found highly significant effects between visualizations with human operators for overall (n = 672, U = 64008, $\rho < 0.001$) and easy decisions (n = 375, U = 19845, $\rho < 0.001$). A moderate significant effect significant effects between visualizations with human operators was found for hard decisions (n = 297, U = 12761, $\rho < 0.01$). Highly significant effects were found between human operators and simulation for the IA visualization overall (U = 34650, ρ

Table 4.25: Selection success rate (%) descriptive statistics per decision difficulty.

	Model	**Decision Difficulty**	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	Overall	75 (43.37)	100 (0/100)
		Easy	81.44 (38.98)	100 (0/100)
		Hard	66.2 (47.47)	100 (0/100)
	M_{2SIM}	Overall	73.69 (19.01)	70 (20/100)
		Easy	77.15 (27.12)	85.71 (0/100)
		Hard	62.05 (27.17)	66.67 (0/100)
Collective	M_2	Overall	88.39 (32.08)	100 (0/100)
		Easy	94.44 (22.97)	100 (0/100)
		Hard	81.41 (39.03)	100 (0/100)
	M_{2SIM}	Overall	74.58 (18.39)	70 (20/100)
		Easy	76.7 (26.87)	83.33 (0/100)
		Hard	64.04 (26.38)	66.67 (0/100)

< 0.001), easy (n = 481, U = 18242, $\rho < 0.001$), and hard decisions (n = 384, U = 13088, $\rho < 0.001$) and for the Collective visualization overall (U = 22162, $\rho < 0.001$), easy (n = 461, U = 10795, $\rho < 0.001$), and hard decisions (n = 392, U = 9449.5, $\rho < 0.001$).

The Spearman correlation analysis revealed a moderate correlation between human-collective team decision time and selection success rate using the IA visualization for easy decisions (r = -0.42, $\rho < 0.001$). Weak correlations were revealed between the human-collective team decision time and selection success rate using the IA visualization for overall decisions (r = -0.27, $\rho < 0.001$) and when using the Collective visualization for overall (r = -0.11, $\rho = 0.05$), easy (r = -0.18, $\rho = 0.02$), and hard decisions (r = 0.18, $\rho = 0.03$). Moderate correlations were revealed between simulation decision time and selection success rate using the IA visualization for overall decisions (r = -0.53, $\rho < 0.001$) and when using the Collective visualization for overall (r = -0.57, $\rho < 0.001$) and easy decisions (r = -0.44, $\rho < 0.001$). Weak correlations were revealed between simulation decision time and selection success rate using the IA visualization for easy (r = -0.35, $\rho < 0.001$) and hard decisions (r = -0.14, $\rho = 0.03$).

SA probe accuracy, which is the percentage of correctly answered SA probes questions used to assess the operator's SA during a trial, results identified that Collective operators had higher SA probe accuracy, regardless of the SA level. Further details about the statistical tests were provided in Chapter 4.1.1.1.

Spearman correlation analyses were conducted to see if any correlations were identified between the weekly hours that participants' used a desktop or laptop, video game proficiency, the mental rotations assessment, and selection success rate. A weak correlation was found between weekly hours participants' used a desktop or laptop and selection success rate for the IA visualization and easy decisions (r = 0.16, $\rho = 0.02$).

The *post-trial performance and understanding* questionnaire results, assessed the operators' understanding of the collective behavior and their ability to chose the best target for each decision, were ranked higher for the Collective visualization compared to the IA. The statistical test details were provided in Chapter 4.1.2.1.

A summary of R_4's results that show the hypotheses with associated significant results is shown in Table 4.26. This summary table is intended to facilitate the discussion.

Table 4.26: A synopsis of R_4's hypothesis (H_8) associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question.

Variable	**Sub-Variable**	**Between Visualization**	**Correlation**	
			IA	Coll.
Decision Time	Overall	$\boldsymbol{H_8}$	H_8	$\boldsymbol{H_8}$
	Easy	$\boldsymbol{H_8}$		$\boldsymbol{H_8}$
	Hard	$\boldsymbol{H_8}$		H_8
Selection Success Rate	Overall	$\boldsymbol{H_8}$	———	
	Easy	$\boldsymbol{H_8}$		
	Hard	$\boldsymbol{H_8}$		
SA Probe Accuracy	SA_O	$\boldsymbol{H_8}$		
	SA_1	$\boldsymbol{H_8}$		
	SA_2	$\boldsymbol{H_8}$		
	SA_3	$\boldsymbol{H_8}$		
Weekly Hours on Desktop or Laptop	SA_1		H_8	
Post-Trial	Understanding	$\boldsymbol{H_8}$	———	

4.1.4.2 Discussion

The analysis suggests that the Collective visualization promoted better human-collective *performance*. H_8 was supported, because the Collective visualization produced higher selection success rate and *SA performance*, as well as higher subjective *performance*. Op-

erators using the Collective visualization were more *efficient* by making significantly faster decisions and having higher selection success rates compared to the IA visualization. Realistic human-collective user scenarios will require high *performance* in short decision *times*, especially in proposed high-risk environments. Longer decision *times* contributed to higher success rates for both visualizations for overall decisions and for easy decisions for the IA visualization; however, faster decision *times* contributed to higher selection success for hard decisions using the Collective visualization. The design of an *effective* human-collective system must enable the human-collective team to fulfill primary objectives, without hindering other metrics, such as decision *time*. Devoting more *time* to ensure high task *performance* is a common trade-off behavior observed in teams. Expedited decisions may have occurred if higher valued targets were more *observable* further away from other objects (less clutter), making them more salient, or if impatient operators were able to *predict* future collective behaviors and influenced collectives more to make decisions faster. The *explainability* of the information pop-up windows aided Collective operators to answer more *SA* probe questions accurately. The Collective visualization enabled operators with different *capabilities* to perform relatively the same, unlike the IA visualization, which found that individuals with more weekly desktop or laptop exposure had higher selection success rates.

The transparency embedded in the Collective visualization promoted the fastest decision times, selection success rates, and *SA performance*. Strategies, such as providing control mechanisms to undo actions that had negative influence on collective behaviors, and providing supplementary information, promoted transparency. *Understanding* what interactions contributed to higher *performance* is necessary to determine what operator strategies are most *effective* and *efficient*, as well as identify what visualization characteristics are being leveraged in order to *perform* the task successfully.

4.1.5 Visualization Analysis Discussion

The research objective was to determine which visualization achieved better transparency, identify what metrics were useful in determining better transparency, and to create design guidance for human-collective systems. The analysis indicated that the Collective visualization provided better transparency, because operators with different individual *capabilities performed* similarly for both the primary and secondary tasks, and the human-collective team *performed* better. The Mental Rotations Assessment, NASA-TLX, and SART were useful individual operator *capability* metrics when determining the influence of the visualization on the operator and can be easily used in other collective evaluations. Post-trial subjective assessments, such as the NASA-TLX and SART; however, will be biased by memory limitations, emphasizing the need for objective metrics that can assess similar *information*. Operator experience (e.g., weekly hours on a desktop or laptop) and expertise (e.g., video game proficiency) can indicate the desired operator knowledge in order to interact with the collective system *effectively*.

The influence of visualization on human operator comprehension and visualization *usability* requires further investigation in order to better *understand* the influence of operator interactions and identify more reliable metrics to assess operator *understanding*. Using correlations between an operator's interactions and *SA performance* can be used to inform designers whether the actions taken by operators aided their responses, but does not necessarily provide insight regarding comprehension. The use of eye-tracking technology can provide improved insight regarding operator comprehension and *usability* by recording where the operator was looking 15 seconds before asking, while asking, and during response to a *SA* probe question. Where operators are looking on the visualization prior to taking action will indicate what types of *information* the opera-

tor was potentially perceiving and comprehending, the difficulty to identify the desired *information* due to visualization clutter, and the duration of *time* devoted to looking at particular *information*. Clutter will greatly impact an operator's ability to perceive and comprehend *information* on the visualization and is an informative metric to use when assessing transparency for human-collective systems.

The reliance on collective and target *information* pop-up windows contributed to the majority of clutter for both visualizations and suggests alterations to design recommendations for future human-collective systems. Providing supplementary *information*, via *information* pop-up windows, was necessary for successful human-collective *performance*; however, other strategies must be implemented to improve the efficacy of the collective icon. Indicating which collective state is most supported, by displaying either U, F, C, or X, instead of the status of all four states, may be more advantageous to the operator. The target a collective is favoring, committed, or executing may also be displayed on the collective icon. For example, if a collective is favoring Target 8, the collective icon can show F8, which stands for Favoring Target 8. Providing the most supported state and target may enable the operator to quickly *understand* what the collective is doing and determine if interventions, or more support is needed to ensure successful decision-making. Showing the predominant collective state; however, requires the operator to remember what U, F, C, and X stand for, which can be mitigated by adding this *information* to the legend. Bolding the predominant state letter on the current collective icon can be used as an alternative design change to improve the operator's *understanding* of the most supported collective state. The perceived mental demand and effort associated with the Collective visualization may decrease with strategies that make the collective state more obvious to the operator.

The dynamic and streamlined behavior of the individual collective entities on the IA visualization may have aided operator *understanding* of collective behavior, mitigating the need to access as many *information* pop-up windows. Displaying all of the individual entities is not ideal as collectives become larger in size. The increased number of entities will contribute to more clutter, potentially hindering the operator's perception and comprehension. Using iconography, such as arrows pointing in the direction of the most supported target, or providing *predicted* hub locations, may improve operator *understanding* for abstract collective visualizations. Further analysis is needed to verify the *effectiveness* of the proposed strategies. The consistent improvement in decision *time* and selection success rate across all decision difficulties suggests that using visualizations that show all the individual collective entities does not contribute to better human-collective *performance*. The Collective visualization enabled better human-collective *performance*, which is valuable as collective systems become more complex, with improved *capabilities* and the utilization of heterogeneous collectives. Presenting individual collective entities may have caused stress, or confusion, and required operators to slow down the collective decision-making process in order to attain higher selection success rates. *SA performance*, selection success rates, and decision *times* were useful metrics that can be used in other collective system evaluations.

Indicating target value through the use of color and opacity in general embedded transparency successfully. Further analysis is required to determine if the entire target icon must represent the target value to be more perceivable, since operators using the Collective visualization, which used half of the target icon to represent the value and the other half to represent support from the highest supporting collective, abandoned the highest value target more frequently. Opacity levels must be validated to ensure the distinction of low-, medium-, and high-valued targets from one another. Reiterating the

task objective, to choose and move each collective to the highest value target, numerous times during training may help mitigate operator mis*understanding*. Improvements can be made to the *information* pop-up windows in order to decrease the number of reissued abandon commands for operators who relied and verified abandonment using the *information* pop-up windows. When an abandon command is issued, the corresponding target *information* pop-up window can immediately report zero support, instead of the actual collective support, which corresponds closer to operators' mental models. A colored bar can be overlaid on the particular collective that abandoned the target as a secondary measure to ensure the operator *understands* the collective status. Operators can focus their attention on collective support values that are not highlighted, in case the current target is the highest value target for another nearby collective. Using metrics, such as the number of times abandoned requests exceed abandoned targets, can be used as an error metric for collective systems.

Visualization transparency for human-collective systems can be achieved via different design strategies and must be assessed holistically by *understanding* how different factors impact transparency and are influenced by transparency. The four secondary research questions assessed categories of transparency factors that contribute to an *effective* system: 1) operator individual *capabilities*, 2) operator comprehension , 3) visualization *usability*, and 4) human-collective team *performance*. An ideal visualization will enable operators with different individual *capabilities* to *perform* relatively the same, promote operator comprehension, be usable, and promote high human-collective *performance*. The Collective visualization enabled operators with different individual *capabilities* to *perform* relatively the same and promoted better human-collective *performance*. The IA visualization enabled operators to perceive collective behaviors and collective support for targets more readily than the Collective visualization, where operators used

the collective and target *information* pop-up windows to affirm these types of collective behaviors. As collective systems grow in complexity (e.g., number of individual agents, heterogeneity), visualizations that show all of the individual collective entities will cause perceptual and comprehension challenges, as well as influence operator actions negatively, because too many individual collective entities will clutter the display. The same advantageous observation (i.e., dynamically seeing collective behaviors and support) from this analysis may not occur with large collectives (> 10000). Abstract collective visualizations may help promote better transparency, than visualizations showing all of the individual collective entities and enable *effective* human-collective teams.

4.2 Model with Visualization Analysis

The primary objective of the within-model and between-visualization analysis was to determine which model and visualization combination achieved better transparency. Four secondary research question's assessed how transparency influenced operators with different individual capabilities, operator comprehension, system usability, and human-collective performance. The subset of direct and indirect factors (Figure 2.2) were assessed in the Model with Visualization Analysis and are identified in Figure 4.7. The hypotheses, metrics, results, and discussions for the Model with Visualization Analysis are presented in Chapters 4.2.1 - 4.2.4, which correspond to research questions R_5 - R_8. The questions and hypotheses were comparable to those in the Visualization Analysis from Chapter 4.1 in order to assess the findings from both analyses. A research question specific representation of the analyzed direct and indirect transparency factors is provided. The Model with Visualization Analysis is concluded with a final discussion that incorporates each respective secondary research question discussion.

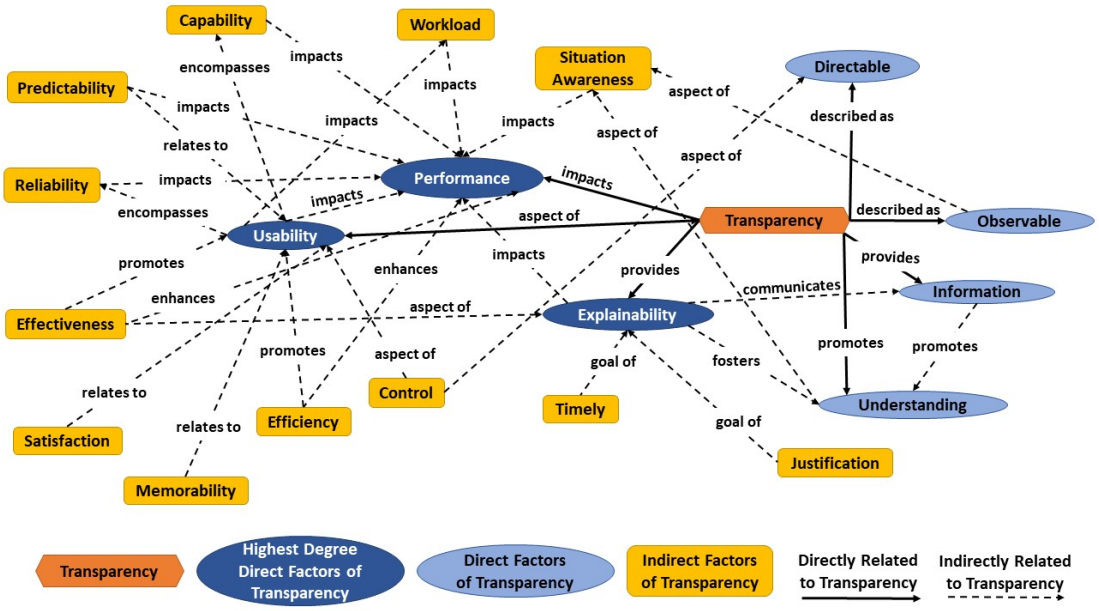


Figure 4.7: The analyzed direct and indirect transparency factors included in the Model with Visualization Analysis.

4.2.1 R_5: System Design Element Influence on Human Operator

Understanding *how the model and visualization influenced the operator*, R_5, is necessary to determine if the system transparency aided operators with different capabilities. The associated objective dependent variables were (1) the operator's ability to influence the collective in order to choose the highest valued target, (2) SA, (3) visualization clutter, (4) the operator's spatial reasoning, and (5) the operator's working memory capacity. The specific direct and indirect transparency factors related to R_5 are identified in Figure 4.8. The relationship between the variables and the hypotheses, as well as the direct and indirect transparency factors, are shown in Table 4.27.

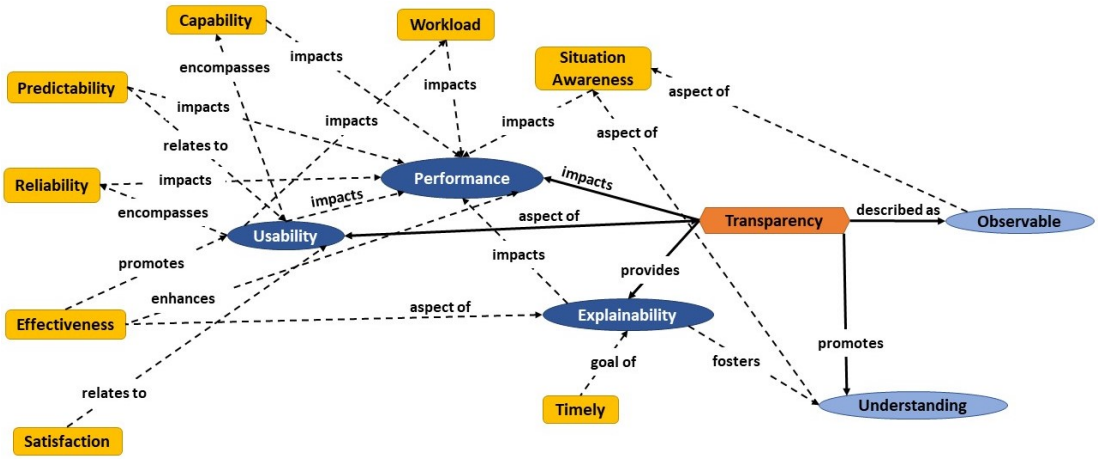


Figure 4.8: R_5 concept map of the assessed direct and indirect transparency factors.

The hypotheses in this chapter and the subsequent result 4.2.2 - 4.2.4 are phrased using the M_2 model with the Collective visualization, because each individual system design element provided the best transparency in their respective evaluations, the Collective evaluation [29] and the IA evaluation of Chapter 4.1 [177]. Operators may have performed differently depending on their individual capabilities; hence, it was hypothesized (H_9) that operators using the M_2 model with the Collective visualization will experience significantly higher SA and lower workload. It was also hypothesized (H_{10}) that operators with different individual capabilities will not perform significantly different using the M_2 model with the Collective visualization. Ideal system design elements will enable operators with different capabilities to perceive, comprehend, and influence collectives relatively the same. Good designs promote higher operator satisfaction. It was hypothesized (H_{11}) that operators using the M_2 model with the Collective visualization will experience significantly less frustration (i.e., higher satisfaction).

Table 4.27: Interaction of system design elements influence on the human operator objective (obj) and subjective (subj) variables (vars), relationship to the hypotheses (H), as well as the associated direct and indirect transparency factors, as shown in Figure 2.2.

		Transparency Factors												
		Direct					**Indirect**							
Obj Vars	**H**	**Explainability**	**Observable**	**Performance**	**Understanding**	**Usability**	**Capability**	**Effectiveness**	**Predictability**	**Reliability**	**SA**	**Satisfaction**	**Timing**	**Workload**
Target Value	H_9					✓								
SA Probe Accuracy	H_9	✓	✓	✓	✓			✓	✓		✓			
Global Clutter	H_9	✓	✓	✓	✓	✓		✓	✓		✓			
Mental Rotations Assessment	H_{10}			✓			✓							
Working Memory Capacity	H_{10}			✓			✓							
Subj Vars														
Weekly Hours on a Desktop or Laptop	H_{10}			✓			✓							
NASA-TLX	H_9, H_{11}	✓		✓			✓					✓	✓	✓
Post-Experiment	H_{10}				✓	✓	✓			✓			✓	

4.2.1.1 Metrics and Results

The *selected target value* is the average of all target's respective values that were selected by the human-collective teams during a trial. The mean (SD) for the selected

target value per decision difficulty (i.e., overall, easy, and hard) are shown in Table 4.28 [177]. IA operators using the M_2 model chose higher valued targets compared to the M_3 model, regardless of the decision difficulty, while Collective operators using the M_3 model chose higher valued targets for overall and hard decisions. The target value median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models for each model with visualization combination are presented in Figure 4.9. IA operators had significantly different selected target values between models for overall and easy decisions, while no differences were found for the Collective operators. Additional Mann-Whitney-Wilcoxon tests were conducted between visualizations and identified moderate significant effects when using the M_3 model for overall (n = 672, U = 63946, $\rho < 0.01$) and highly significant effects for hard decisions (n = 276, U = 12058, $\rho < 0.001$). Collective operators were able to influence the collective to choose higher valued targets compared to those using the IA visualization.

Table 4.28: Selected target value mean (SD) by decision difficulty (Dec Diff), where the maximum possible value was 100 and the minimum possible value was 67.

	Dec Diff	IA	Collective
M_2	Overall	90.29 (7.11)	92.05 (5.08)
	Easy	90.21 (7.29)	92.09 (5.54)
	Hard	90.4 (6.88)	92 (4.5)
M_3	Overall	89.52 (8.05)	92.22 (4.34)
	Easy	90.3 (7.31)	91.73 (4.59)
	Hard	88.39 (8.93)	92.92 (3.88)

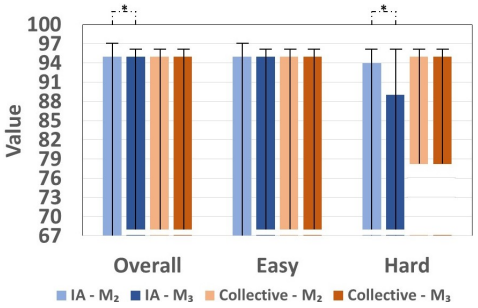


Figure 4.9: Target value median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty with significance ($\rho < 0.001$ - ***, $\rho < 0.01$ - **, and $\rho < 0.05$ - *) between models.

SA probe accuracy was the percentage of correctly answered SA probes questions [29]. Five SA_1 questions determined the operator's ability to perceive information about the collectives and targets, four SA_2 questions determined the operator's comprehension of information, and three SA_3 questions were related to the operator estimating the collectives' future state. Examples of SA probe questions were provided in Chapter 4.1.1.1. The SA probe accuracy mean (SD) are shown in Table 4.29 [57, 171, 177]. Operators from both evaluations using the M_2 model, when compared to M_3 model, had higher SA_3, while the IA operators had higher SA_2, and the Collective operators had higher SA_O. The SA probe accuracy median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.10. Significant differences between models were found for IA operators answering SA_1 probe questions and for Collective operators answering SA_3 probe questions. Additional Mann-Whitney-Wilcoxon tests were conducted between visualizations (n = 56) and identified highly significant effects using the M_2 model for SA_O (U = 702, $\rho < 0.001$) and SA_1 (U = 714.5, $\rho < 0.001$); and moderately significant effects for SA_2 (U = 572.5, $\rho < 0.01$) and SA_3 (U = 554, $\rho < 0.01$).

Table 4.29: SA probe accuracy (%) mean (SD) by SA level.

	Level	IA	Collective)
M_2	SA_O	65.3 (18.87)	89.88 (10.96)
	SA_1	58.57 (23.05)	91.67 (11.11)
	SA_2	72.32 (21.88)	88.39 (14.6)
	SA_3	65.48 (34.52)	89.88 (20.46)
M_3	SA_O	68.15 (16.36)	87.2 (10.75)
	SA_1	80 (19.63)	94.05 (13)
	SA_2	65.18 (28.33)	91.43 (12.68)
	SA_3	52.38 (27.86)	76.79 (16.57)

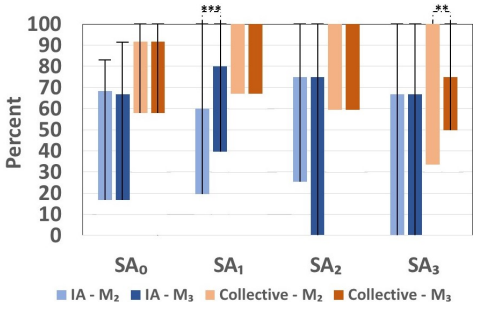


Figure 4.10: SA probe accuracy median (min/max) and Mann-Whitney-Wilcoxin test by SA level with significance ($\rho < 0.001$ - ***, $\rho < 0.01$ - **, and $\rho < 0.05$ - *) between models.

Highly significant effects between visualizations were found using the M_3 model for SA_O (U = 657.5, $\rho < 0.001$), SA_2 (U = 648, $\rho < 0.001$), and SA_3 (U = 645.5, $\rho < 0.001$). A moderately significant effect between visualizations was found using the M_3 model for SA_1 (U = 564, $\rho < 0.01$). Operators using the Collective visualization had higher SA probe accuracy in general.

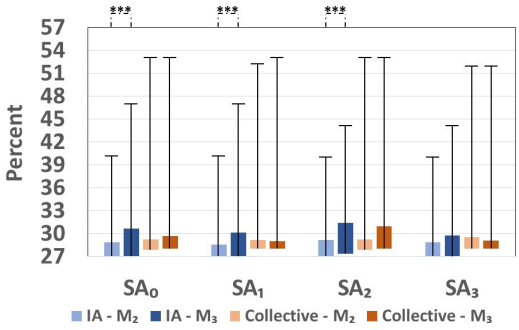
Table 4.30: Global clutter mean (SD) percentage 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	**SA Level**	**IA**	**Collective**
M_2	Before	SA_O	30.2 (3.06)	31.37 (4.97)
		SA_1	29.88 (2.8)	31.38 (5)
		SA_2	30.41 (3.05)	31.25 (5.09)
		SA_3	30.45 (3.45)	31.56 (4.76)
	Asking	SA_O	30.25 (3.13)	31.43 (5.13)
		SA_1	29.95 (2.91)	31.24 (5.26)
		SA_2	30.41 (3.12)	31.52 (5.2)
		SA_3	30.52 (3.49)	31.69 (4.78)
	Responding	SA_O	30.09 (3.02)	31.41 (5.15)
		SA_1	29.83 (2.81)	31.43 (5.43)
		SA_2	30.22 (3)	31.34 (5.08)
		SA_3	30.37 (3.38)	31.49 (4.66)
M_3	Before	SA_O	31.26 (3.41)	31.76 (5.23)
		SA_1	31.2 (3.48)	31.51 (5.05)
		SA_2	31.78 (3.4)	32.11 (5.21)
		SA_3	30.68 (3.24)	31.51 (5.51)
	Asking	SA_O	31.49 (3.59)	31.7 (5.23)
		SA_1	31.6 (3.74)	31.15 (5.05)
		SA_2	31.83 (3.54)	32.33 (5.52)
		SA_3	30.82 (3.34)	31.4 (4.9)
	Responding	SA_O	31.16 (3.36)	31.7 (5.27)
		SA_1	31.25 (3.4)	31.24 (5.12)
		SA_2	31.49 (3.41)	32.25 (5.56)
		SA_3	30.56 (3.2)	31.37 (4.93)

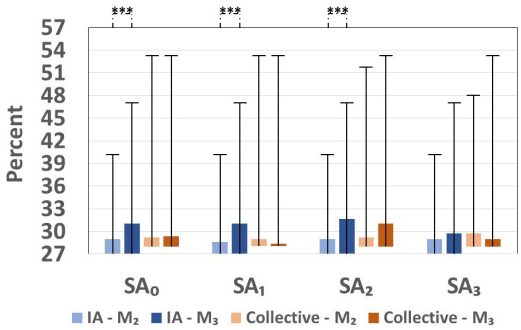
Global clutter percentages were analyzed for each SA probe question and was the percentage of area obstructed by all objects displayed on the computer displays. It was calculated using Equation 4.2 from Chapter 4.1.1.1. The global clutter mean (SD) percentage 15 seconds before asking, while being asked, and during response to a SA probe question are shown in Table 4.30 [177]. IA operators who used the M_2 model had lower global clutter percentages compared to when they used the M_3 model. Collective operators in general had lower global clutter percentages using the M_2 model. SA_3 at all timings and SA_1 while being asked a SA probe question were lower when Collective operators used the M_3 model. The global clutter percentage median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.11. Significant differences between models were found for IA operators at all timings for SA_O, SA_1, and SA_2 probe questions, while only one significant difference between models was identified for Collective operators during response to SA_2 probe questions.

Additional between visualizations Mann-Whitney-Wilcoxin tests were conducted. Significant differences between visualizations occurred when using the M_2 model. A highly significant effect between visualizations was found when responding to a SA probe question for SA_O (n = 670, U = 64442, $\rho < 0.001$). Moderate significant effects between visualizations were found for SA_O 15 seconds before asking (U = 64188, $\rho < 0.01$) and while being asked a SA probe question (U = 63728, $\rho < 0.01$). Significant effects between visualizations were found 15 seconds before asking a SA probe question for SA_1 (n = 294, U = 12487, $\rho = 0.02$) and SA_3 (n = 152, U = 3445.5, $\rho = 0.03$); while being asked a SA probe question for SA_1 (U = 12301, $\rho = 0.03$) and SA_3 (U = 3452, $\rho = 0.05$); and during the response to a SA probe question for SA_1 (U = 12216, $\rho = 0.04$). Correlations between the global clutter percentage and SA probe accuracy were only revealed when using the Collective visualization 15 seconds before asking a SA probe question. The

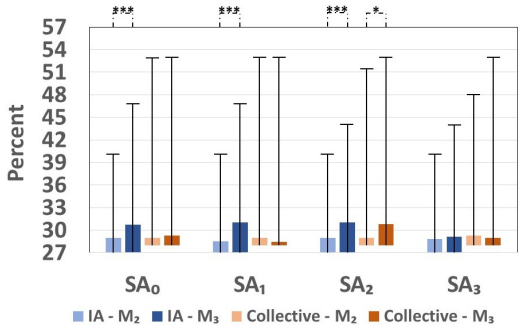
Spearman correlation analysis revealed a moderate correlation with the M_3 model for SA_3 (r = 0.45, $\rho < 0.001$), and weak correlations with the M_2 model for SA_1 (r = 0.16, ρ = 0.05) and with the M_3 model for SA_O (r = 0.2, $\rho < 0.001$). The IA visualization had lower global clutter percentages in general compared to the Collective visualization. Collective operators using the M_3 model; however, had lower global clutter while being asked and during response to a SA_1 probe question.



(a) 15 seconds before asking a SA probe question.



(b) While being asked a SA probe question.



(c) During response to a SA probe question.

Figure 4.11: Global clutter percentage median (min/max) and Mann-Whitney-Wilcoxin test by SA level between models a) 15 seconds before asking, b) while being asked, and c) during response to a SA probe question.

There were no significant effects between visualizations for operator spatial reasoning, based on the Mental Rotations Assessment [172]. Correlations between the Mental

Rotations Assessment and SA probe accuracy only existed for the IA visualization. The Spearman correlation analysis revealed weak correlations with the M_2 model for SA_O (r = 0.17, $\rho < 0.01$), SA_1 (r = 0.18, $\rho = 0.03$), and SA_2 (r = 0.27, $\rho < 0.01$). Weak correlations were revealed with the M_3 model for SA_O (r = 0.15, $\rho < 0.01$), SA_1 (r = 0.19, $\rho = 0.03$), and SA_2 (r = 0.18, $\rho = 0.05$). A moderate correlation existed between Working Memory Capacity, which assessed operator higher-order cognitive task abilities [173], and SA probe accuracy for the IA visualization with the M_2 model for SA_3 (r = 0.45, $\rho < 0.001$). Weak correlations existed with the M_2 model for SA_O (r = 0.23, $\rho < 0.001$) and SA_1 (r = 0.17, $\rho = 0.04$), and with the M_3 model for SA_O (r = 0.14, $\rho = 0.01$). The Mann-Whitney-Wilcoxon tests identified no significant effects between visualizations for the weekly hours spent using a desktop or laptop. Weak correlations were found between weekly hours using a desktop or laptop and SA probe accuracy when using the M_2 model with the IA visualization for SA_O (r = 0.12, $\rho = 0.04$) and SA_1 (r = 0.21, $\rho = 0.01$), and with the Collective visualization for SA_2 (r = 0.21, $\rho = 0.02$).

The *NASA-TLX* assessed the six workload subscales and the weighted overall workload [174]. The mean (SD) for the NASA-TLX overall workload and imposed demands are presented in Table 4.31 [57, 177]. IA operators using the M_2 model had lower physical demand and effort when compared to M_3, while those using the Collective visualization had lower physical demand, effort, and frustration with the M_2 model. The NASA-TLX median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.12. IA operators had significantly different rankings between models for physical demand and frustration, while mental demand was significantly different between models for Collective operators. Additional between visualizations Mann-Whitney-Wilcoxin tests (n = 56) identified a significant effect when using the M_2 model for mental demand (U = 515, $\rho = 0.04$) and a highly

significant effect for performance (U = 159.5, $\rho < 0.001$). Significant effects were found between visualizations using the M_3 model for overall workload (U = 266.5, $\rho = 0.04$), performance (U = 242.5, $\rho = 0.01$), and frustration (U = 511, $\rho = 0.05$), as well as a highly significant effect for physical demand (U = 208, $\rho < 0.001$). The Collective visualization imposed a lower overall workload, had lower physical and temporal demands, and caused less frustration compared to the IA visualization.

Table 4.31: NASA-TLX mean (SD).

	TLX	IA	Collective
M_2	Overall	62.14 (14.81)	57.06 (16.47)
	Mental	19.25 (8.8)	23.58 (6.34)
	Physical	1.68 (3.32)	0.46 (1.17)
	Temp.	11.75 (8.24)	10.94 (7.67)
	Perfor.	10.69 (5.87)	5.1 (4.7)
	Effort	11.35 (6.68)	12.32 (6.36)
	Frus.	7.43 (8.36)	4.65 (6.84)
M_3	Overall	60.38 (16.5)	50.63 (17.56)
	Mental	18.32 (9.4)	16.54 (9.19)
	Physical	6.11 (10.27)	1.81 (6.01)
	Temp.	8.85 (7.3)	7.49 (6.63)
	Perfor.	9.08 (6.7)	5.15 (4.79)
	Effort	14.25 (8.06)	12.5 (5.17)
	Frus.	3.77 (5.92)	7.14 (8.31)

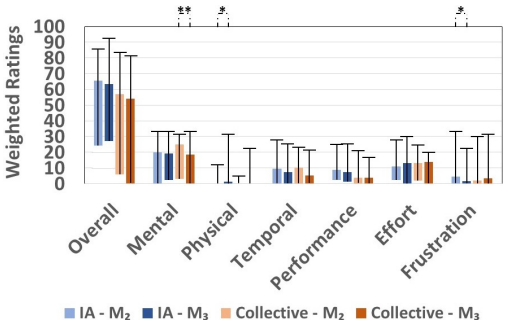


Figure 4.12: NASA-TLX median (min/max) and Mann-Whitney-Wilcoxin test between models.

The post-experiment questionnaire assessed the collective's *responsiveness* to requests, the participants' *ability* to choose the highest valued target, and their *understanding* of the collective behavior, from best (1) to worst (2 for the IA evaluation and 3 for the Collective evaluation). The post-experiment questionnaire mean (SD) are presented in Table 4.32 [29]. The best collective responsiveness as well as operator ability and understanding occurred when IA operators used the M_2 model versus the M_3 model. Collective operators ranked the collective's responsiveness highest with the M_3 model,

while operator ability and understanding were highest with the M_2 model. The post-experiment questionnaire median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.13. System responsiveness, operator ability, and understanding were ranked significantly different between models for IA operators, while Collective operators ranked system responsiveness and operator understanding significantly different.

Table 4.32: Post-experiment responsiveness, ability, and understanding model ranking mean (SD) (1-best, 2-worst for IA evaluation and 3-worst for Collective evaluation).

	Metric	IA	Collective
M_2	Responsive.	1.64 (0.49)	1.5 (0.51)
	Ability	1.86 (0.36)	2 (1.02)
	Understand.	1.79 (0.42)	2.5 (0.51)
M_3	Responsive.	1.36 (0.49)	3 (0)
	Ability	1.14 (0.36)	2 (0)
	Understand.	1.21 (0.42)	1 (0)

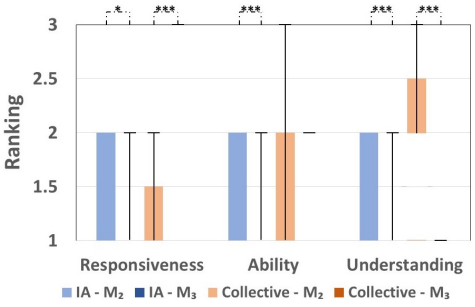


Figure 4.13: Post-experiment responsiveness, ability, and understanding model ranking median (min/max) and Mann-Whitney-Wilcoxin test between models. The ranking was from 1-best to either 2-worst for the IA evaluation, or 3-worst for the Collective evaluation.

A summary of R_5's results that show the hypotheses with associated significant results is shown in Table 4.33. This summary table is intended to facilitate the discussion.

4.2.1.2 Discussion

Relationships to the transparency factors provided in Table 4.27 are emphasized using italics. The analysis of how the model and visualization influenced operators with dif-

Table 4.33: A synopsis of R_5's hypotheses associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question.

Variable	**Sub-Variable**	**Within Model**		**Between Visualization**		**Correlation**			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Target Value	Overall	H_9			H_9				
	Hard	H_9			H_9				
SA Probe Accuracy	SA_O			***H*9**	H_9				
	SA_1	H_9		***H*9**	H_9				
	SA_2			***H*9**	H_9				
	SA_3			***H*9**	H_9				
Global Clutter Percentage	SA_O	H_9 -AT		H_9 -AT					H_9 -B
	SA_1	H_9 -AT		H_9 -AT				H_9 -B	
	SA_2	H_9 -AT	H_9 $-D$	H_9 -B,W					
	SA_3								H_9 -B
Mental Rotations Assessment	SA_O					H_{10}	H_{10}		
	SA_1					H_{10}	H_{10}		
	SA_2						H_{10}		
Working Memory Capacity	SA_O					H_{10}	H_{10}		
	SA_1					H_{10}			
	SA_3					H_{10}			
Weekly Hours on Desktop or Laptop	SA_O					H_{10}			
	SA_1					H_{10}			
	SA_2							H_{10}	
NASA-TLX	Overall				H_9				
	Mental		H_9	H_9					
	Physical	H_9			H_9				
	Perfor.			***H*9**	H_9				
	Frus.	H_9, H_{11}			H_9, H_{11}				

Variable	Sub-Variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Post-Experiment	Ability	H_9		———					

ferent individual *capabilities* suggests that the M_2 model promoted transparency as *effectively* as the M_3 model, while the Collective visualization promoted better transparency compared to the IA visualization. H_9, which hypothesized that operators using the M_2 model and Collective visualization will experience significantly higher *SA* and lower *workload*, was not supported. *SA performance* (i.e., accuracy) varied across the SA levels depending on the model and *workload* varied across the *workload* subscales depending on the model and visualization. The M_2 model was *effective* at enabling operators to more accurately *predict* future collective behaviors, while the M_3 model enabled better *observability* of the collectives' behaviors. Better *predictability* may have occurred, because the M_2 model aligned with the operators expectations: that the model was designed to choose the highest value target. *Predictability* of future collective states may have also improved due to the visualization. Favoring entities in the IA visualization created streamlines between hubs and targets, which may have directed the operator's attention to particular targets. The M_3 model may have promoted better perception of the collectives' behaviors, because the operator was required to direct those behaviors in order to achieve the task. Operator *workload* was alleviated by the M_2 model by requiring less operator *capabilities*, such as physical demand and effort, as well as promoting higher *performance*, which was expected since operator influence was not required in order to make decisions. The M_3 model alleviated operator *workload* by also requiring less operator *capabilities*, such as mental demands, and improving *satisfaction*

(i.e., less frustration) by mitigating temporal (i.e., *timing*) demands. More operator control of the decision-making process, such making decision quickly or more slowly, may have contributed to these lower *workload* subscales.

Transparency embedded into the Collective visualization partially supported H_9, because it promoted higher *SA performance* via the color-coded icons and outlines, state information identified on the collective icon, information provided in the pop-up windows, as well as feedback provided in the Collective Assignments and System Messages areas. Collective operators encountered more clutter; however, due to the long duration of time the target information pop-up windows were visible. The increased clutter has both positive and negative implications for transparency. Clutter, from an *usability* perspective, is not ideal if operators are unable to perform their tasks *effectively*. The Collective operators, who had higher clutter were able to answer more *SA* probe questions accurately, which suggests that the operators were not hindered by the clutter and *performed* better. The dependence on the visible target information pop-up windows may have been caused by the type of *SA* probe questions asked. Thirteen of twenty-four *SA* probe questions relied on information provided in the target information-pop up windows. An example question, such as "What collectives are investigating Target 3?", required using the target information pop-up window, if Target 3 was in range of multiple collectives. The operator was able to identify which collectives were within range of a particular target by left-clicking on the respective collective; however, target information pop-up windows were needed in order to see the numeric collective support values from a specific collective, or multiple collectives. Experimental design modifications can ensure a more even distribution of *SA* probe questions that rely on other information, such as the icons, system messages, or collective assignments versus information pop-up windows. Target icon design modifications that indicate

which collectives support a particular target may improve *explainability*, *reliability*, and increase the reliance on the target icon instead of the information pop-up window.

The Collective visualization partially supported H_9 by requiring less operator *capabilities*, such as physical demands, and improving *satisfaction* (i.e., less frustration) by mitigating temporal (i.e., *timing*) demands. Not visualizing entities may have reduced operator stress, because the rate of a collective's state change was not easily perceived. The need or desire to influence collective behaviors may not have been as apparent, which attributed to lower physical demand and frustration. Higher operator mental demand when using the M_2 model and Collective visualization may have occurred if collective behaviors, or state changes, were not *observable* and required more interactions to deduce what was happening, such as accessing information pop-up windows.

H_{10}, which hypothesized that operators with different individual *capabilities* did not *perform* significantly different using the M_2 model and the Collective visualization, was partially supported. Individuals with different spatial reasoning and working memory capacity *capabilities performed* relatively the same. Operators who had a higher level of computer knowledge; however, had a better *understanding* of the collective behaviors. This finding was anticipated considering the computer simulation environment. Further investigations are needed to identify what particular aspects of computer knowledge attribute to better *understanding*.

Collective operators using the M_2 model were more *satisfied* (i.e., less frustration), which supported H_{11}. Dis*satisfaction* transpires when the system is not transparent and prohibits the operator from *understanding* what is currently happening, or the interface appears visually noisy due to clutter [163]. A more autonomous model, one with more decision-making *capabilities*, and an abstract collective visualization may mitigate dis*satisfaction*. More metrics, such as the Questionnaire for User Interface Satisfaction

[178], are needed to properly assess how the transparency embedded in the models and visualizations influence operator *satisfaction*.

The transparency embedded in the Collective visualization in general supported operators with individual differences better than the IA visualization. The M_2 model; however, did not support all operators. More computer experience, for example, aided operator *SA performance*. Mitigating the need for operators to have particular *capability* levels is desired in order to design *effective* human-collective systems. Higher *SA performance* also varied between the models, which suggests system design changes must be considered in order to improve the perception, comprehension, and projection of future collective behaviors when using the M_2 model. *Usability* considerations need to identify the ideal amount of operator influence in the decision-making process in order to alleviate *workload* (e.g., mental demand) and promote better *SA*.

4.2.2 R_6: System Design Element Promotion of Operator Comprehension

The explainability direct transparency factor was explored in R_6, which was interested in determining whether the transparency embedded in *the model and visualization promoted operator comprehension*. The associated objective dependent variables were (1) SA, (2) collective and target left- or right-clicks, (3) collective and target observations, (4) interventions, (5) the percentage of times the highest value target was abandoned, and (6) whether the information pop-up window was open when a target was abandoned. The specific direct and indirect transparency factors related to R_6 are identified in Figure 4.14. The relationship between the variables and the corresponding hypotheses, as

well as the direct and indirect transparency factors, are identified in Table 4.34.

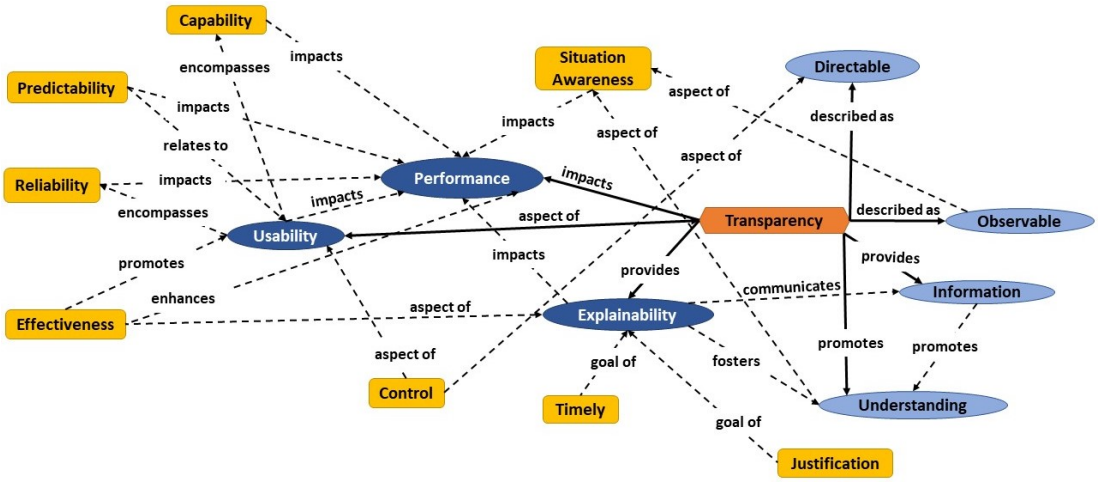


Figure 4.14: R_6 concept map of the assessed direct and indirect transparency factors.

Models designed to aid operators to fulfill a best-of-n decision-making task can help mitigate workload by reducing repetitive interactions, ensure task progress in case an operator becomes distracted or is attending to a different collective's decision process, and allow more time to establish SA and understanding. Using display principles may help improve understanding by providing legible, clear, concise, organized, easily accessible, and consistent information. It was hypothesized (H_{12}) that operators will have a better understanding of the M_2 model with the information provided by the Collective visualization. Appropriate expectations of the model's capabilities and contributions towards a goal, as well as providing information redundantly via icons, colors, messages, and the information pop-up windows can aid operator comprehension and justify future actions. It was hypothesized (H_{13}) that operators using the M_2 model with the Collective visualization were able to accurately justify actions.

Table 4.34: Interaction of system design elements promotion of human operator comprehension objective (obj) and subjective (subj) variables (vars), relationship to the hypotheses (H), as well as the associated direct and indirect transparency factors, are presented in Figure 2.2.

		Transparency Factors													
		Direct							**Indirect**						
Obj Vars	**H**	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Capability**	**Control**	**Effectiveness**	**Justification**	**Predictability**	**Reliability**	**SA**	**Timing**
SA Probe Accur	H_{12}				✓	✓	✓				✓		✓		✓	
Coll Left-Clicks	H_{13}		✓					✓			✓	✓				
Target Right-Clicks by SA Level	H_{13}		✓					✓			✓	✓				
Coll Observ	H_{13}		✓	✓				✓			✓	✓				
Target Observ	H_{12}		✓				✓	✓								
Coll Right-Clicks	H_{13}		✓	✓				✓			✓	✓				
Target Right-Clicks per Dec	H_{13}		✓	✓				✓			✓	✓				

		Transparency Factors														
		Direct							Indirect							
Obj Vars	**H**	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Capability**	**Control**	**Effectiveness**	**Justification**	**Predictability**	**Reliability**	**SA**	**Timing**
Interv	H_{12}	✓	✓				✓	✓		✓	✓		✓	✓		
High Value Target Aban	H_{12}, H_{13}	✓	✓				✓	✓		✓	✓	✓				
Aban Target Info Wind Open	H_{13}		✓					✓			✓	✓				
Subj Vars																
Post-Trial Perf Under	H_{12}		✓			✓	✓		✓							
Post-Exper	H_{12}						✓	✓	✓					✓		✓

4.2.2.1 Metrics and Results

The results of *SA probe accuracy*, which is the percentage of correctly answered SA probes questions used to assess the operator's SA during a trial, identified that IA and Collective operators using the M_2 model had higher SA_3 compared to the M_3 model, while the IA operators had higher SA_2 and the Collective operators had higher SA_O. Operators using the Collective visualization had higher SA probe accuracy, regardless of the SA level, compared to the IA visualization. Further details regarding the statisti-

cal tests were provided in the Metrics and Results Chapter 4.2.1.1.

Collective left-clicks identified all targets within range of a collective and was the first click required to issue a command. The number of collective left-clicks mean (SD) 15 seconds before asking, while being asked, and during response to a SA probe question are shown in Table 4.35 [177]. The M_2 model in general had fewer collective left-clicks compared to the M_3 model. Collective operators using the M_3 model while being asked

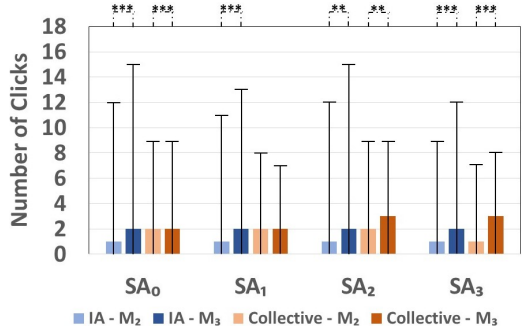
Table 4.35: Collective left-clicks mean (SD) 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	**SA Level**	**IA**	**Collective**
M_2	Before	SA_O	1.64 (1.84)	1.95 (1.57)
		SA_1	1.53 (1.75)	1.88 (1.47)
		SA_2	1.78 (1.9)	2.13 (1.68)
		SA_3	1.65 (1.92)	1.83 (1.61)
	Asking	SA_O	0.49 (0.76)	0.69 (0.88)
		SA_1	0.3 (0.6)	0.51 (0.79)
		SA_2	0.42 (0.77)	0.91 (0.89)
		SA_3	0.33 (0.61)	0.73 (0.96)
	Responding	SA_O	1.68 (1.79)	1.52 (1.21)
		SA_1	1.14 (1.46)	1.32 (1.02)
		SA_2	1.46 (1.8)	1.57 (1.21)
		SA_3	1.53 (1.98)	1.89 (1.48)
M_3	Before	SA_O	2.48 (2.28)	2.58 (1.76)
		SA_1	2.42 (2.15)	2.27 (1.73)
		SA_2	2.45 (2.33)	2.71 (1.7)
		SA_3	2.61 (2.43)	2.79 (1.85)
	Asking	SA_O	0.63 (0.82)	0.85 (0.83)
		SA_1	0.46 (0.65)	0.88 (0.87)
		SA_2	0.91 (0.89)	0.83 (0.8)
		SA_3	0.52 (0.88)	0.85 (0.86)
	Responding	SA_O	2.02 (1.81)	1.97 (1.39)
		SA_1	1.78 (1.7)	1.83 (1.25)
		SA_2	2.21 (1.78)	2.04 (1.43)
		SA_3	2.16 (1.98)	2.05 (1.5)

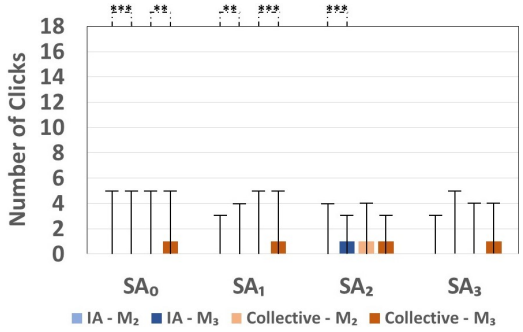
a SA probe question had fewer collective left-clicks for SA_2. The number of collective left-clicks median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.15. IA operators had significantly different collective left-clicks between models for SA_O, SA_1, and SA_2 at all timings, as well as SA_3 15 seconds before asking and during response to a SA probe question. Significantly different collective left-clicks between models were found 15 seconds before asking a SA probe question for SA_O, SA_2, and SA_3, while being asked a SA probe question for SA_O and SA_1, and during response to a SA probe question for SA_O, SA_1, and SA_2.

Additional between visualizations Mann-Whitney-Wilcoxon tests found highly significant effects when using the M_2 model 15 seconds before asking a SA probe question for SA_O (n = 664, U = 64213, $\rho < 0.001$), a moderate significant effect for SA_1 (n = 290, U = 12534, $\rho < 0.01$), and a significant effect for SA_2 (n = 223, U = 7210.5, $\rho = 0.04$). Highly significant effects between visualizations using the M_2 model while being asked a SA probe question were found for SA_O (U = 67670, $\rho < 0.001$), and SA_2 (U = 8317 $\rho < 0.001$), as were moderately significant effects for SA_1 (U = 12043, $\rho < 0.01$), and SA_3 (n = 151, U = 3472, $\rho < 0.01$). A highly significant effect between visualizations using the M_2 model during response to a SA probe question was found for SA_O (U = 64710, $\rho < 0.001$), a moderate significant effect for SA_1 (U = 12414, $\rho < 0.01$), and a significant effect for SA_3 (U = 3489, $\rho = 0.01$). A significant effect between visualizations when using the M_3 model 15 seconds before asking a SA probe question was found for SA_O (n = 665, U = 60696, $\rho = 0.03$). Highly significant effects between visualizations using the M_3 model while being asked a SA probe question were found for SA_O (U = 64376, $\rho < 0.001$), SA_1 (n = 251, U = 9959.5, $\rho < 0.001$), as well as a moderate significant effect for SA_3 (n = 162, U = 4114, $\rho < 0.01$). Correlations between the collective left-clicks and SA probe accuracy were only revealed using the M_3 model. The Spearman correlation analysis

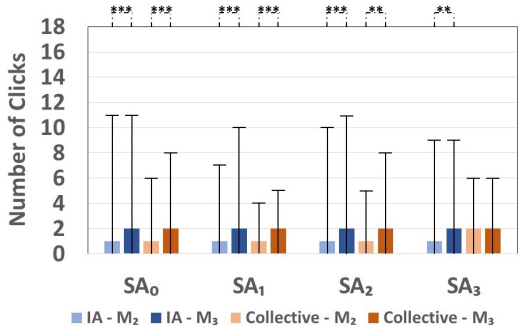
revealed weak correlations with the IA visualization for SA_3 15 seconds before asking (r = -0.26, ρ = 0.02) and while being asked a SA probe question (r = -0.33, $\rho < 0.01$). Weak correlations were also revealed with the Collective visualization while being asked a SA probe question for SA_O (r = 0.13, ρ = 0.02) and SA_1 (r = 0.22, ρ = 0.02). The IA visualization had fewer collective left-clicks in general compared to the Collective visualization. Collective operators who used the M_2 model during response to a SA probe question had fewer left-clicks for SA_O, and with the M_3 model for all SA levels.



(a) 15 seconds before asking a SA probe question.



(b) While being asked a SA probe question.



(c) During response to a SA probe question.

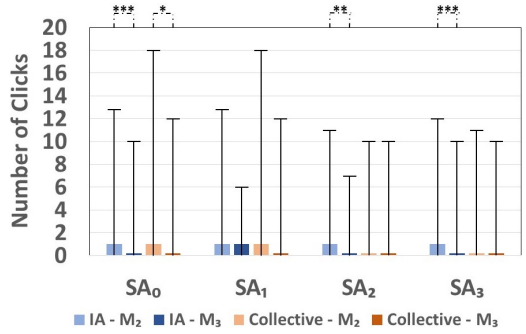
Figure 4.15: Collective left-clicks median (min/max) and Mann-Whitney-Wilcoxin test by SA level between models a) 15 seconds before asking, b) while being asked, and c) during response to a SA probe question.

Target right-clicks allowed the operator to access target information pop-up windows that provided each collective's percentage of support for a respective target. The number of target right-clicks mean (SD) 15 seconds before asking, while being asked, and during response to a SA probe question are presented in Table 4.36 [177]. The M_2 model in general had fewer target right-clicks for both visualizations. Collective operators who used the M_3 model while being asked a SA probe question had fewer target right-

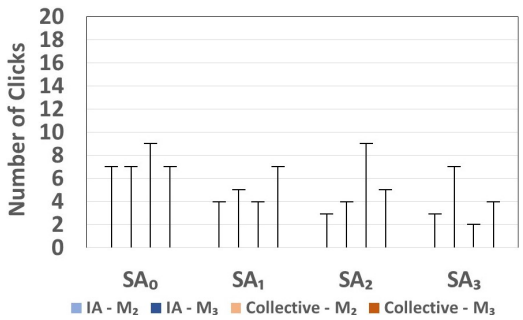
Table 4.36: Target right-clicks mean (SD) 15 seconds before asking, while being asked, and during response to SA probe question by SA level.

	Timing	**SA Level**	**IA**	**Collective**
M_2	Before	SA_O	1.68 (2.38)	1.52 (2.41)
		SA_1	1.92 (2.62)	1.79 (2.71)
		SA_2	1.28 (1.91)	1.17 (1.94)
		SA_3	1.8 (2.5)	1.49 (2.35)
	Asking	SA_O	0.37 (0.79)	0.5 (1)
		SA_1	0.44 (0.74)	0.49 (0.86)
		SA_2	0.31 (0.67)	0.55 (1.31)
		SA_3	0.37 (0.75)	0.44 (0.69)
	Responding	SA_O	1.07 (1.77)	0.99 (1.7)
		SA_1	1.11 (1.69)	1.01 (1.74)
		SA_2	1.1 (1.75)	0.84 (1.44)
		SA_3	1.68 (2.24)	1.21 (1.98)
M_3	Before	SA_O	1.04 (1.68)	1.17 (2)
		SA_1	1.34 (1.54)	1.44 (2.44)
		SA_2	0.79 (1.53)	0.96 (1.63)
		SA_3	0.88 (2.03)	1.15 (1.88)
	Asking	SA_O	0.36 (0.86)	0.42 (0.95)
		SA_1	0.42 (0.82)	0.38 (0.96)
		SA_2	0.29 (0.79)	0.44 (0.98)
		SA_3	0.35 (1)	0.45 (0.9)
	Responding	SA_O	0.89 (1.64)	0.72 (1.3)
		SA_1	0.91 (1.69)	0.67 (1.49)
		SA_2	0.8 (1.68)	0.72 (1.17)
		SA_3	0.98 (1.52)	0.8 (1.24)

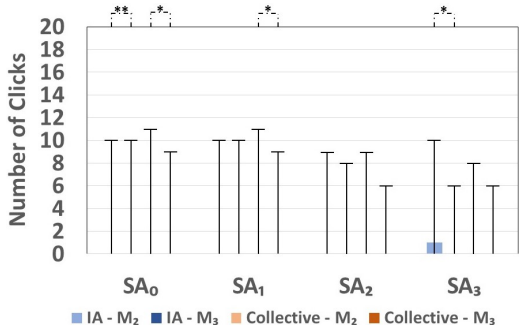
clicks for SA_3. The number of target right-clicks median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.16. IA operators had significantly different collective left-clicks between models 15 seconds before asking a SA probe question for SA_O, SA_2, and SA_3, as well as during response to a SA probe question for SA_O and SA_3. Significantly different collective left-clicks between models were found 15 seconds before asking a SA probe question for SA_O and during response to a SA probe question for SA_O and SA_1.



(a) 15 seconds before asking a SA probe question.



(b) While being asked a SA probe question.



(c) During response to a SA probe question.

Figure 4.16: Target right-clicks median (min/max) and Mann-Whitney-Wilcoxin test by SA level between models a) 15 seconds before asking, b) while being asked, and c) during response to a SA probe question.

No significant effects between visualizations were found. Collective operators using the M_2 model had fewer target right-clicks for all SA levels, 15 seconds before asking and during response to a SA probe question compared to the IA operators. Fewer target right-clicks, 15 seconds before asking and while being asked a SA probe question, occurred when IA operators used the M_3 model compared to the Collective visualization. The Spearman correlation analysis revealed weak correlations between the number of target right-clicks and SA probe accuracy using the M_2 model with the IA visualization 15 seconds before asking a SA probe question for SA_O (r = 0.17, $\rho < 0.01$) and SA_2 (r = 0.37, $\rho < 0.001$). Weak correlations were found using the M_3 model with the IA visualization 15 seconds before asking a SA probe question for SA_O (r = 0.11, $\rho = 0.04$), SA_1 (r = 0.2, $\rho = 0.02$), and with the Collective visualization for SA_1 15 seconds before asking (r = -0.24, $\rho = 0.01$) and while being asked a SA probe question (r = -0.21, $\rho = 0.03$).

Collective observations were collective left-clicks that only identified targets within range of a collective (i.e., white borders indicated that the individual collective entities were investigating the target, while yellow indicated no investigation) and whether the targets had been abandoned (i.e., red borders). The percentage of collective left-clicks associated with collective observations mean (SD) by decision difficulty are shown in Table 4.37 [57]. IA operators using the M_3 model had fewer collective observations compared to the M_2 model, while Collective operators had fewer collective observations when using the M_2 model. The collective observations median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.17. IA operators had significantly different collective observations between models for all decision difficulties, while Collective operators had significantly different collective observations between models for easy decisions. Additional between visualizations Mann-Whitney-Wilcoxon tests identified a moderate significant effect when using

the M_2 model for overall (n = 672, U = 61152, $\rho < 0.01$) and a significant effect for easy decisions (n = 374, U = 19008, $\rho = 0.05$). Highly significant effects between visualizations when using the M_3 model were found for overall (U = 73920, $\rho < 0.001$), easy (n = 396, U = 25587, $\rho < 0.001$), and hard decisions (n = 276, U = 12520, $\rho < 0.001$). The IA visualization had fewer collective observations compared to the Collective visualization.

Table 4.37: Collective observations (%) mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	77.68 (41.7)	86.01 (34.74)
	Easy	71.13 (45.43)	80 (40.11)
	Hard	86.62 (34.16)	92.95 (25.68)
M_3	Overall	59.23 (49.21)	90.18 (29.8)
	Easy	57.79 (49.51)	88.32 (32.19)
	Hard	61.31 (48.88)	92.81 (25.93)

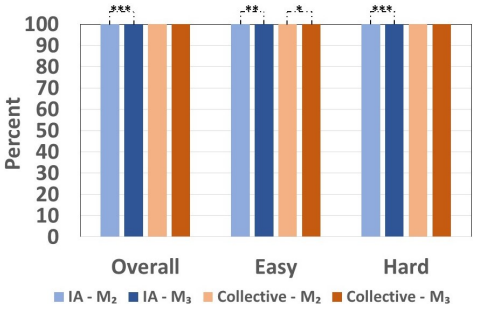


Figure 4.17: Collective observations median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

Target observations represent the subset of target left-clicks not associated with issuing a command. The percentage mean (SD) for target left-clicks that were target observations by decision difficulty are shown in Table 4.38 [57]. The M_2 model with the Collective visualization had fewer target observations, regardless of decision difficulty. The target observations median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.18. IA operators had significantly different target observations between models for overall decisions, while Collective operators had significantly different target observations between models for all decision difficulties. Additional between visualizations Mann-Whitney-Wilcoxon tests identified highly significant effects when using the M_2 model for overall (n = 672, U = 35280, $\rho < 0.001$), easy (n = 374, U = 10886, $\rho < 0.001$), and hard decisions (n = 298, U = 6910,

$\rho < 0.001$). Highly significant effects between visualizations when using the M_3 model were also found for overall (U = 41664, $\rho < 0.001$), easy (n = 396, U = 15053, $\rho < 0.001$), and hard decisions (n = 276, U = 6615, $\rho < 0.001$).

Table 4.38: Target observations (%) mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	60.12 (49.04)	22.62 (41.9)
	Easy	58.76 (49.35)	21.11 (40.92)
	Hard	61.97 (48.72)	24.36 (43.06)
M_3	Overall	67.26 (47)	41.07 (49.27)
	Easy	64.32 (48.03)	41.12 (49.33)
	Hard	71.53 (45.29)	41.01 (49.36)

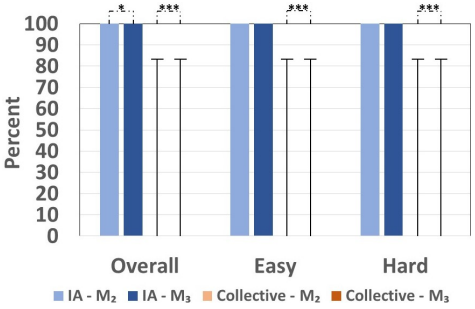


Figure 4.18: Target observations median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

Collective right-clicks allowed the operator to open or close collective information pop-up windows, which provided the number of individual collective entities in each decision-making state. Operators may have used the information to justify issuing commands. The number of collective right-clicks per decision was only assessed for the IA evaluation, because the Collective evaluation did not record which particular collective pop-up window was opened or closed. The number of collective right-clicks mean (SD) per decision difficulty are presented in Table 4.39. The M_3 model had fewer collective right-clicks compared to the M_2 model, regardless of decision difficulty. The collective right-clicks median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.19. Significantly different collective right-clicks between models were found for overall and hard decisions.

Table 4.39: Collective right-clicks per decision mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA
M_2	Overall	1.55 (2.25)
	Easy	1.16 (1.87)
	Hard	2.09 (2.6)
M_3	Overall	0.88 (2.2)
	Easy	0.87 (2.54)
	Hard	0.89 (1.57)

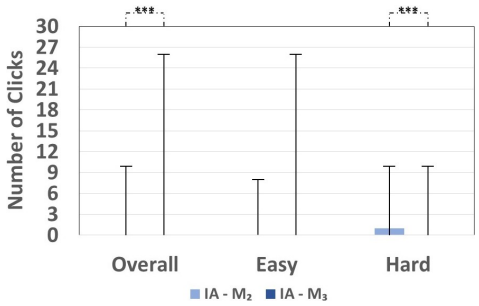


Figure 4.19: Collective right-clicks median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

Target right-clicks allowed the operator to open or close target information pop-up windows, which provided the percentage of support each collective had for a respective target. The target support information may have also been used to justify issuing commands, such as increasing or decreasing support from particular collectives. The mean (SD) for the number of target right-clicks by decision difficulty are presented in Table 4.40. IA operators using the M_2 model had fewer target right-clicks, while Collective operators had fewer target right-clicks using the M_3 model. The target right-clicks

Table 4.40: Target right-clicks per decision mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	3.54 (4.18)	3.09 (3.56)
	Easy	2.64 (3.14)	2.61 (2.87)
	Hard	4.77 (5.03)	3.64 (4.17)
M_3	Overall	3.75 (5.38)	3.04 (3.49)
	Easy	3.8 (5.82)	2.95 (3.46)
	Hard	3.67 (4.69)	3.15 (3.54)

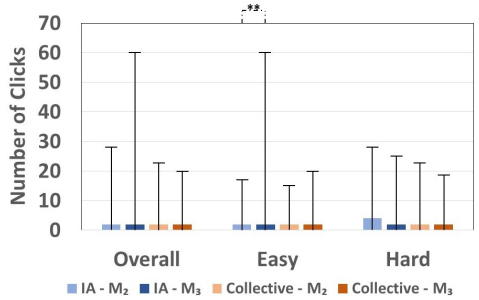


Figure 4.20: Target right-clicks median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.20. IA operators had significantly different target right-clicks between models for easy decisions, while no differences were found for the Collective operators. The Collective visualization had fewer target right-clicks compared to the IA visualization; however, no significant effects between visualizations were found.

Interventions occurred when the operator abandoned a target with greater than 10% collective support. Abandoning low-value targets was a desired intervention. Interventions were assessed per participant, due to the inability to associate an intervention to a decision, and the descriptive statistics are shown in Table 4.41 [57]. The M_2 model with the IA visualization had fewer interventions. The Mann-Whitney-Wilcoxon tests found a significant effect between models for the IA visualization (n = 56, U = 270.5, ρ = 0.04). No significant effects between visualizations were found.

Table 4.41: Interventions (abandoned targets with 10% support) per participant descriptive statistics.

	Model	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	1.5 (2.03)	0.5 (0/7)
	M_3	3.75 (4.27	3 (0/17)
Collective	M_2	2.21 (1.99)	1.5 (0/7)
	M_3	5 (5.11)	3.5 (0/18)

The abandon command discontinued a collective's investigation of a particular target. Ideally lower valued targets were abandoned, since the objective was to aid each collective in selecting and moving to the highest valued target. The percentage of times the *highest value target was abandoned* per participant mean (SD) are presented in Table 4.42 [177]. Operators using the M_3 model abandoned the highest value target less frequently compared to the M_2 model. The highest value target abandoned median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are pre-

sented in Figure 4.21. IA operators had significantly different highest value target abandoned percentages between models for easy decisions, while Collective operators had significant differences highest value target abandoned percentages between models for overall decisions. Operators using the IA visualization abandoned the highest value target less frequently compared to those using the Collective visualization; however, no significant effects were found between the visualizations.

Table 4.42: Highest value target abandoned (%) mean (SD) per participant by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	32.36 (29.53)	43.6 (31.94)
	Easy	31.2 (27.17)	33.25 (35.96)
	Hard	42.1 (40.53)	48.72 (36.85)
M_3	Overall	18.56 (18.38)	21.04 (21.19)
	Easy	11.35 (20.82)	11.64 (14.26)
	Hard	22.47 (11.89)	28.09 (22.39)

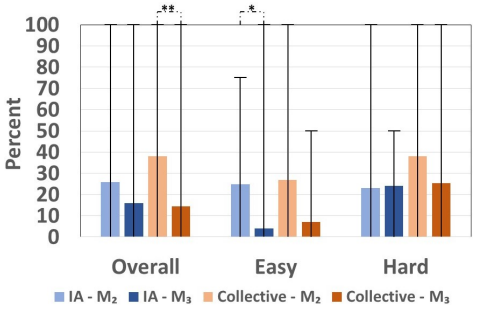


Figure 4.21: Highest value target abandoned median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The percentage of times an *abandoned target information pop-up window was open* per participant was evaluated and the mean (SD) are presented in Table 4.43 [177]. Operators using the M_3 model had fewer abandoned target information pop-up windows open compared to the M_2 model. The abandoned target information pop-up window open median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.22, but no significant effects between models were found. Additional between visualizations Mann-Whitney-Wilcoxon tests identified significant effects using the M_3 model for overall (n = 49, U = 414.5, ρ = 0.02) and easy decisions (n = 45, U = 352, ρ = 0.02). Fewer abandoned target information pop-up windows were open when using the IA visualization compared to the Collective visualization.

Table 4.43: Abandoned target information pop-up window open (%) mean (SD) per participant by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	23.86 (31.43)	33.8 (34.9)
	Easy	22.2 (30.95)	30.7 (37.85)
	Hard	28.7 (37.89)	36.08 (40.87)
M_3	Overall	8.48 (15.6)	26.96 (35.48)
	Easy	9.17 (16.13)	28.18 (34.02)
	Hard	8.65 (18.5)	25.18 (38.69)

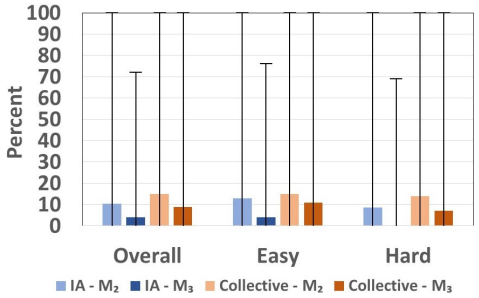


Figure 4.22: Abandoned target information pop-up window open median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The post-trial questionnaire assessed the participants' *understanding of collective behavior*, never (1) to always (7), and their *ability* to choose the best target per decision, never (1) to always (7). The post-trial questionnaire mean (SD) are shown in Table 4.44 [29, 177]. The performance and understanding rankings were higher for Collective operators using the M_3 model. The post-trial performance and understanding median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models

Table 4.44: Post-trial performance and understanding model ranking mean (SD) (1-low, 7-high).

	Metric	IA	Collective
M_2	Performance	5.25 (1.69)	5.54 (1.29)
	Understand.	4.89 (1.75)	5.82 (1.16)
M_3	Performance	5.57 (1.43)	5.75 (1.43)
	Understand.	5.93 (1.02)	5.93 (1.46)

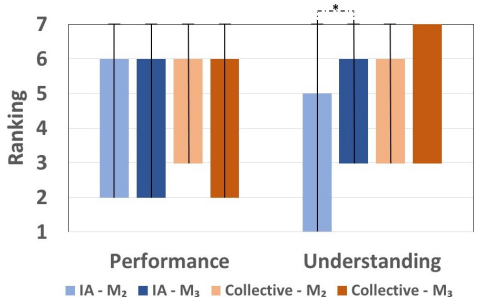


Figure 4.23: Post-trial performance and understanding model ranking median (min/max) and Mann-Whitney-Wilcoxin test between models.

are shown in Figure 4.23. IA operators ranked understanding significantly different between models. Additional between visualizations Mann-Whitney-Wilcoxon tests found a significant effect for understanding using the M_2 model (n = 56, U = 513, ρ = 0.04).

The post-experiment questionnaire assessed the collective's *responsiveness* to requests, the participants' *ability* to choose the highest valued target, and their *understanding* of the collective behavior. IA operators who used the M_2 model had the best collective responsiveness, operator ability, and understanding versus the M_3 model. Collective operators ranked the collective's responsiveness highest using the M_3 model, while operator ability and understanding were highest using the M_2 model. Details regarding the statistical tests were provided in the Metrics and Results Chapter 4.2.1.1.

A summary of R_6's results that show the hypotheses with associated significant results is shown in Table 4.45. This summary table is intended to facilitate the discussion.

Table 4.45: A synopsis of R_6's hypotheses associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question.

Variable	**Sub-Variable**	**Within Model**		**Between Visualization**		**Correlation**			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
SA Probe Accuracy	SA_O			$\boldsymbol{H_{12}}$	H_{12}				
	SA_1	H_{12}		$\boldsymbol{H_{12}}$	H_{12}				
	SA_2			$\boldsymbol{H_{12}}$	H_{12}				
	SA_3			$\boldsymbol{H_{12}}$	H_{12}				

Variable	Sub-variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Collective Left-Clicks	SA_O	H_{13} -AT	$\boldsymbol{H_{13}}$ $\boldsymbol{-AT}$	H_{13} -AT	H_{13} -B,W				H_{13} -W
	SA_1	H_{13} -AT	H_{13} -W,D	H_{13} -AT	H_{13} -W				H_{13} -W
	SA_2	H_{13} -AT	H_{13} -B,D	H_{13}, -B,W					
	SA_3	H_{13} -B,D	H_{13} -B	H_{13} -W,D	H_{13} -W		H_{13} -B,W		
Target Right-Clicks by SA Level	SA_O	H_{13} -B,D	H_{13} -B,D			H_{13} -B	H_{13} -B		H_{13} -B
	SA_1		H_{13} -D				H_{13} -B		H_{13} -B,W
	SA_2	H_{13} -B				H_{13} -B			
	SA_3	H_{13} -B,D							
Collective Observations	Overall	H_{13}		H_{13}	H_{13}				
	Easy	H_{13}	H_{13}	H_{13}	H_{13}				
	Hard	H_{13}			H_{13}				
Target Observations	Overall	H_{12}	$\boldsymbol{H_{12}}$	$\boldsymbol{H_{12}}$	H_{12}				
	Easy		$\boldsymbol{H_{12}}$	$\boldsymbol{H_{12}}$	H_{12}				
	Hard		$\boldsymbol{H_{12}}$	$\boldsymbol{H_{12}}$	H_{12}				
Collective Right-Clicks	Overall	H_{13}							
	Hard	H_{13}							
Target Right-Clicks per Decision	Easy	H_{13}							
Interventions		H_{12}							
Highest Value Target Abandoned	Overall		H_{12}, H_{13}						
	Easy	H_{12}, H_{13}							

Variable	Sub-variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Abandoned Target Info. Window Open	Overall				H_{13}				
	Easy				H_{13}				
Post-Trial	Under.	H_{12}		H_{12}					
Post-Experiment	Under.	H_{12}	H_{12}						

4.2.2.2 Discussion

The analysis of how the model and visualization promoted operator comprehension (i.e., the operator's *capability* of *understanding*) suggests that the M_3 model promoted transparency more *effectively* than the M_2 model, while both visualizations had their respective advantages and disadvantages. Operators using the M_2 model had fewer undesired interactions, such as target observations (i.e., extra clicks that did not contribute to the task) and interventions. Fewer undesired interactions may have occurred, because the M_2 model was designed to fulfill the best-of-n decision-making task with or without operator influence, which *effectively* balanced *control* between the collectives and operator, whereas the M_3 model relied on operator influence (*directability*) in order to make a decision. More undesirable interactions, such as target observations, resulted in better task *performance* for operators using the M_3 model, which suggests that some interactions deemed undesirable for one model may be advantageous for another. Target observations may have occurred due to poor interface and visualization *usability*. Operators who issued commands first selected the desired command, then selected the desired collective and target, and clicked on the commit button to complete a request. Reissuing the same command required re-selecting the target and clicking on the com-

mit button. More target observations may have occurred if operators forgot to re-select the target when reissuing the same commands. Design improvements, such as leaving the target selected, may help decrease target observations.

H_{12}, which hypothesized that operators will have a better *understanding* of the M_2 model, was not supported, because operators using the M_2 model abandoned the highest value target more frequently. The operators may have become overloaded supervising the four collectives simultaneously, especially if they were distracted by the secondary task and were momentarily out-of-the-loop. The interface's 10 Hz update rate (i.e., *timing*) may have negatively impacted the operator's *capability* to *understand* what the collectives were doing and planned (e.g., *predictability*) to do. Introducing *timing* delays to the display may afford operators more *time* to *understanding* the current situation; however, task completion will be prolonged, which is undesired in missions that require fast system responses. Providing *predictive* collective behaviors instead of *timing* delays may help mitigate the *time* required for an operator to reenter back into-the-loop.

The highest value target was abandoned more frequently when using the Collective visualization. The target value may not have been *observable* enough (i.e., salient) to distinguish it from other potential targets, which did not support H_{12}. Further investigations are required to determine if the target value must use the entire collective hub icon area, similar to the IA visualization, in order to be more recognizable, and to establish what levels of obscurity are needed in order to ensure that target values are *reliably* distinguishable from one another. Making distinctions clearer, such as using integers compared to letters, to identify collectives versus targets, may improve visualization *explainability* and mitigate mistakes when operators confused the roman numeral identifiers with the integer identifiers. IA operators experienced this mistake frequently, which may have contributed to lowering their *understanding*. Ensuring that identifiers

are unique and distinct will improve the *effectiveness* of the *SA* probe questions.

The use of target borders (collective observations), *information* pop-up windows (target right-clicks), and target value, were assessed to determine if operators used this *information* to *justify* actions *reliably* (i.e., accurately). Collective operators using the M_2 model made better decisions with fewer collective observations and more target-right clicks. *Understanding* which collectives supported targets, by seeing numerical percentages, was more valuable compared to outlines indicating which targets were within a collective's range. H_{13}, which hypothesized that operators using the M_2 model and the Collective visualization were able to *justify* actions accurately, was not supported. Collective operators who issued more collective left-clicks while being asked a SA probe question had better perception when using the M_3 model. IA operators who issued more target right-clicks 15 seconds before asking a SA probe question had better comprehension when using the M_2 model. The interactions of both operators were accurate and *justified*; however, the model and visualization combination did not support the hypothesis. Collective left-clicks can improve perception of targets in range of a particular collective and are attributed with issued commands, which require perception, comprehension, and projection. Target right-clicks provide more *information* about collective support for a particular target, which may improve *understanding*.

Lower *SA performance* may have occurred if operators were in the middle of an interaction when the *SA* probe question was posed, while higher *SA performance* may have occurred because the operators anticipated when a *SA* probe question was going to be asked and took preventative actions, such as opening or closing *information* windows. Operators using target *information* pop-up windows to verify that a target was abandoned by a collective may have been confused if the reported target support was greater than zero. There were instances during the trial when a few individual entities

became lost, as the collective hub transitioned to a new location, and they did not move with the hub. The lost entities may have continued to explore a now abandoned target, because they never received the abandon target message, which occurred inside of the hub. The operators, as a result, may have reissued additional abandon commands in an attempt to reduce the collective support to zero, although only one abandon command was needed. Strategies improving *explainability*, such as reporting zero percent support when an abandon command is issued and identifying how many individual entities have been lost, may help mitigate erroneous repeated abandon command behavior and improve *understanding*. IA operators may have also experienced confusion if they saw individual collective entities still travelling to an abandoned target. Not displaying lost entities after a specific period of *time* once a collective hub has moved to a new location may also reduce the number of reissued abandon commands. Further analysis using eye-tracking technology may provide more *reliable* metrics to determine operator comprehension by identifying exactly where an operator is focusing their attention.

The transparency embedded in the M_2 model and Collective visualization combination did not support the operator's *capability* to *understand* (i.e., comprehension) the collectives' behaviors the best. The M_3 model provided better operator comprehension, because operators were more involved in the decision-making process. More interactions, even if some were undesirable, contributed to better *understanding* and task *performance*. Strategies to increase operator involvement, without taking complete *control* over the decision-making process, when using the M_2 model must be considered to improve it's *effectiveness*. Design improvements, such as increasing *explainability* by identifying how many individual entities became lost during a hub transition to a new location, can help mitigate abandoning the highest value target, which occurred most frequently for Collective operators using the M_2 model. *Understanding* why particu-

lar interactions occurred for specific model and visualization combinations, and what aspects contributed to those interactions, can help aid designers to improving the transparency embedded in the M_2 model and Collective visualization.

4.2.3 R_7: System Design Element Usability

Understanding *which model and visualization promoted better usability*, R_7, can determine which system design elements promote transparency in human-collective systems.

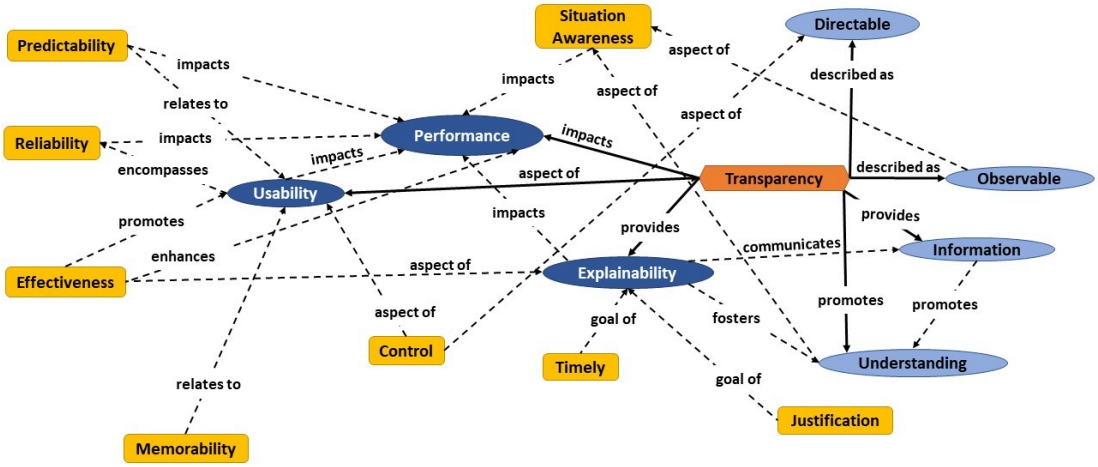


Figure 4.24: R_7 concept map of the assessed direct and indirect transparency factors.

The objective dependent variables were (1) visualization clutter, (2) Euclidean distance, (3) whether an operator was in the middle of an action and completed that action when asked a SA probe question, (4) issued commands, (5) collective and target right-clicks, (6) metrics associated with abandoned targets, (7) the time between the commit state and issued decide command, and (8) metrics associated with information pop-up windows. The specific direct and indirect transparency factors related to R_7 are identi-

fied in Figure 4.24. The relationship between the variables and the hypotheses, as well as the direct and indirect transparency factors are shown in Table 4.46.

The goal of usability is to design systems that are effective, efficient, easy to learn, and are memorable [163]. It was hypothesized (H_{14}) that the M_2 model with the Collective visualization will promote better usability by being more predictable and explainable. Providing information that is explainable may aid comprehension, while predictable information may expedite operator actions. An ideal system will not require constant interaction to perform well; therefore, it was hypothesized (H_{15}) that operators using the M_2 model with the Collective visualization will require fewer interactions.

Table 4.46: Interaction of system design elements usability objective (obj) and subjective (subj) variables (vars), relationship to the hypotheses (H), as well as the associated direct and indirect transparency factors, are presented in Figure 2.2.

		Transparency Factors													
		Direct							**Indirect**						
Obj Vars	**H**	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Control**	**Effectiveness**	**Justification**	**Memorability**	**Predictability**	**Reliability**	**SA**	**Timing**
Global Clutter	H_{14}				✓	✓	✓	✓		✓			✓		✓	
Euclid Dist Betw SA Probe Inter and Clicks	H_{15}							✓		✓						

		Transparency Factors														
		Direct							Indirect							
Obj Vars	**H**	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Control**	**Effectiveness**	**Justification**	**Memorability**	**Predictability**	**Reliability**	**SA**	**Timing**
Midd of Action Dur SA Probe	H_{15}	✓			✓		✓	✓	✓	✓			✓		✓	✓
Comp Interr SA Probe Action	H_{14}, H_{15}	✓	✓		✓		✓	✓	✓	✓		✓	✓		✓	
Invest Comm	H_{15}	✓				✓		✓	✓	✓						
Aban Comm	H_{15}	✓				✓		✓	✓	✓						
Decide Comm	H_{14}, H_{15}	✓				✓		✓	✓	✓			✓			
Coll Right-Clicks	H_{15}			✓		✓		✓		✓						
Target Right-Clicks per Dec	H_{15}			✓		✓		✓		✓						
High Value Target Aban	H_{14}		✓				✓	✓		✓	✓					

		Transparency Factors													
		Direct							Indirect						
Obj Vars	**H**	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Control**	**Effectiveness**	**Justification**	**Memorability**	**Predictability**	**Reliability**	**SA**	**Timing**
Aban Target Info Wind Open	H_{14}		✓					✓		✓	✓					
Aban Req Exced Aban Target	H_{14}		✓				✓	✓		✓						
Time Comm State and Issued Decide Comm	H_{14}				✓		✓	✓		✓			✓		✓	✓
Freq of Access Target Info Wind	H_{14}, H_{15}		✓	✓				✓		✓						
Time Target Info Wind Open	H_{14}							✓		✓				✓		✓

		Transparency Factors													
		Direct							Indirect						
Obj Vars	**H**	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Control**	**Effectiveness**	**Justification**	**Memorability**	**Predictability**	**Reliability**	**SA**	**Timing**
Time Dec Coll Info Wind Open	H_{14}							✓		✓						✓
Time Dec Target Info Wind Open	H_{14}							✓		✓						✓
Subj Vars																
Post-Trial Comm Effect	H_{14}							✓		✓				✓		
Post-Exper	H_{14}						✓	✓						✓		✓

4.2.3.1 Metrics and Results

System features were available to the operators in order to aid task completion. The IA visualization had lower *global clutter percentages*, which was the percentage of visualization area obstructed by all displayed objects. IA operators using the M_2 model had lower global clutter percentages compared to the M_3 model. Collective operators in general had lower global clutter percentages using the M_2 model. The IA visualization

had lower global clutter percentages in general compared to the Collective visualization. The statistical test details were provided in Chapter 4.2.1.1.

The Euclidean *distance between the SA probe interest and where the operator was interacting* with the visualization indicated where operators focused their attention. The Euclidean distance was calculated using the method previously mentioned in Chapter 4.1.3.1. The Euclidean distance between SA probe interest and clicks mean (SD) 15 sec-

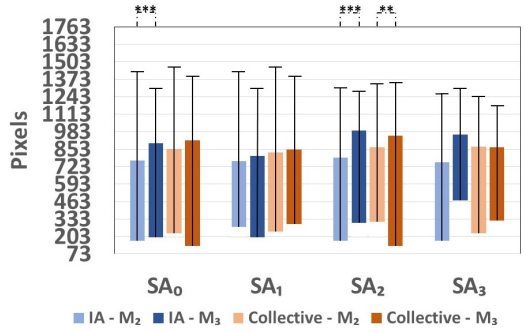
Table 4.47: Euclidean distance between SA probe interest and clicks mean (SD) 15 seconds before asking, while being asked, and during response to SA probe by SA level.

	Timing	**SA Level**	**IA**	**Collective**
M_2	Before	SA_O	767.1 (262.5)	820.7 (255.67)
		SA_1	759.5 (251.64)	825.6 (264.1)
		SA_2	768.9 (282.07)	812.9 (234.94)
		SA_3	783.4 (262.89)	821.6 (271.03)
	Asking	SA_O	758.44 (291.48)	851.4 (293.91)
		SA_1	754.4 (284.65)	845.5 (282.53)
		SA_2	768.4 (316.09)	879.5 (299.93)
		SA_3	753.7 (275.04)	823.5 (314.47)
	Responding	SA_O	764.24 (298.84)	827.7 (273.83)
		SA_1	760.9 (297.14)	827.9 (279.21)
		SA_2	774.6 (319.08)	845.2 (275.55)
		SA_3	757.71 (278.14)	799.7 (261.1)
M_3	Before	SA_O	868.3 (239.4)	845.1 (258.07)
		SA_1	814.4 (225.39)	789.9 (261.93)
		SA_2	925.2 (243.31)	896.9 (241.67)
		SA_3	907.8 (238.96)	805.9 (277.15)
	Asking	SA_O	862.3 (254.68)	860.22 (266.91)
		SA_1	808.4 (250.01)	846.4 (272.23)
		SA_2	931.6 (249.36)	933.5 (252.66)
		SA_3	865.7 (241.6)	759 (248.65)
	Responding	SA_O	865 (262.27)	837 (263.75)
		SA_1	816.7 (254.45)	802.7 (270.79)
		SA_2	928.3 (264.62)	901.6 (238.64)
		SA_3	860.4 (248.47)	755.2 (274.87)

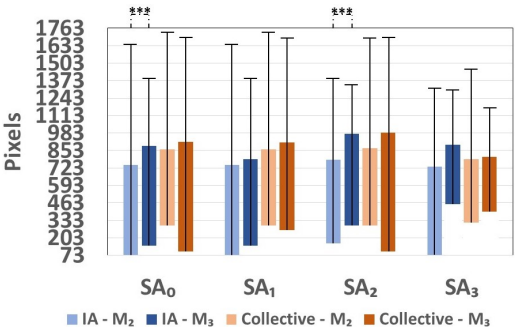
onds before asking, while being asked, and during response to a SA probe question are shown in Table 4.47 [177]. Operators from both visualizations using the M_2 model in general had shorter Euclidean distances compared to the M_3 model. Collective operators using the M_3 model; however, had shorter Euclidean distances at all timings for SA_3 and 15 seconds before asking and during response to a SA probe question for SA_1.

The Euclidean distance between SA probe interest and operator clicks median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.25. IA operators had significantly different Euclidean distances between the SA probe interest and their current interaction between models 15 seconds before asking, while being asked, and during response to a SA probe question for SA_O and SA_2. A significant difference for this metric between models occurred for Collective operators 15 seconds before asking a SA_2 probe question. Additional between visualizations Mann-Whitney-Wilcoxon tests identified significant effects when using the M_2 model 15 seconds before asking a SA probe question for SA_O (n = 557, U = 43303, ρ = 0.02) and SA_1 (n = 273, U = 10577, ρ = 0.05). A moderate significant effect between visualizations when using the M_2 model while being asked a SA probe question was found for SA_O (n = 464, U = 31052, $\rho < 0.01$) as was a significant effect for SA_1 (n = 229, U = 7645, ρ = 0.01). A significant effect between visualizations using the M_2 model during response to a SA probe question was also found for SA_O (n = 499, U = 35029, ρ = 0.02). Shorter Euclidean distances occurred when IA operators used the M_2 model compared to the Collective visualization, while Collective operators had shorter Euclidean distances when using the M_3 model. The Spearman correlation analysis revealed a weak correlation between the Euclidean distance of the SA probe's interest and the operators' current click and SA probe accuracy when using the M_2 model with the IA visualization 15 seconds before asking a SA probe question for SA_1 (r = -0.18, ρ = 0.04). Weak correlations were revealed

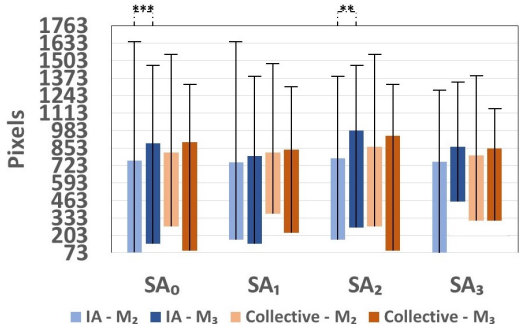
when using the M_3 model with the Collective visualization for SA_O while being asked (r = 0.14, ρ = 0.04) and during response to a SA probe question (r = 0.16, ρ = 0.01).



(a) 15 seconds before asking a SA probe question.



(b) While being asked a SA probe question.



(c) During response to a SA probe question.

Figure 4.25: Euclidean distance between SA probe interest and clicks median (min/max) and Mann-Whitney-Wilcoxin test by SA level between models a) 15 seconds before asking, b) while being asked, and c) during response to a SA probe question.

The percentage of times an operator was in the *middle of an action during a SA probe* question identified how often operators were interrupted by the secondary task. Distracted operators may have needed more time to focus their attention on the SA probe question, or may have prioritized their current interaction over answering the SA probe question immediately, or at all. Understanding how distractions may have negatively influenced operator behavior is needed to design the system to promote effective human-

collective interactions. The percentage of times an operator was in the middle of an action during a SA probe question mean (SD) are shown in Table 4.48.

Table 4.48: Middle of an action during SA probe (%) mean (SD) by SA level.

	Level	IA	Collective
M_2	SA_O	13.47 (34.19)	47.02 (49.99)
	SA_1	10.71 (31.04)	46.1 (50.01)
	SA_2	13.39 (34.21)	46.43 (50.1)
	SA_3	18.29 (38.9)	50 (50.36)
M_3	SA_O	27.68 (44.81)	66.67 (47.21)
	SA_1	28.37 (45.24)	66.96 (47.25)
	SA_2	26.79 (44.48)	69.29 (46.3)
	SA_3	27.71 (45.03)	61.9 (48.85)

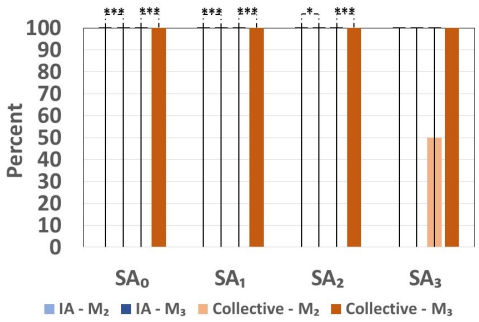


Figure 4.26: The percentage of times a participant was in the middle of an action during a SA probe question median (min/max) and Mann-Whitney-Wilcoxin test by SA level between models.

Operators using the M_2 model were interrupted less by the SA probe question compared to those using the M_3 model irrespective of the visualization. The percentage of times operators were in the middle of an action during a SA probe question median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are identified in Figure 4.26. The percentage of times operators from both evaluations were in the middle of an action during a SA probe question was significantly different between models for SA_O, SA_1, and SA_2. Additional between visualizations Mann-Whitney-Wilcoxon tests identified highly significant effects when using the M_2 model for SA_O (n = 670, U = 74938, $\rho < 0.001$), SA_1 (n = 294, U = 14595, $\rho < 0.001$), SA_2 (n = 224, U = 8344, $\rho < 0.001$), and SA_3 (n = 152, U = 3780, $\rho < 0.001$). Highly significant effects between visualizations using the M_3 model were found for SA_O (n = 672, U = 78456, $\rho < 0.001$), SA_1 (n = 253, U = 10944, $\rho < 0.001$), SA_2 (n = 252, U = 11172, $\rho < 0.001$),

and SA_3 (n = 167, U = 4678, $\rho < 0.001$). IA operators were interrupted less frequently by SA probe questions. The Spearman correlation analysis revealed weak correlations between the middle of an action during a SA probe question and SA probe accuracy for the IA visualization using the M_2 model for SA_1 (r = -0.22, $\rho < 0.01$), as well as the M_3 model for SA_2 (r = 0.19, $\rho = 0.05$) and SA_3 (r = -0.33, $\rho < 0.01$). A weak correlation was revealed for the Collective visualization using the M_2 model for SA_3 (r = 0.24, $\rho = 0.05$).

Table 4.49: Completed interrupted SA probe action (%) mean (SD) by SA level.

	Level	IA	Collective
M_2	SA_O	98.8 (10.89)	98.81 (10.86)
	SA_1	100 (0)	98.7 (11.36)
	SA_2	99.11 (9.45)	98.21 (13.3)
	SA_3	96.34 (18.89)	100 (0)
M_3	SA_O	100 (0)	98.51 (12.13)
	SA_1	100 (0)	98.21 (13.3)
	SA_2	100 (0)	98.57 (11.91)
	SA_3	100 (0)	98.81 (10.91)

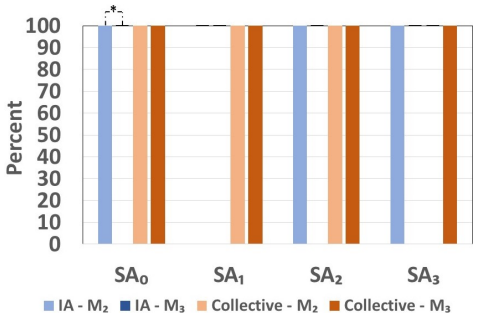


Figure 4.27: Completed interrupted SA probe action median (min/max) and Mann-Whitney-Wilcoxin test by SA level between models.

The percentage of times a participant *completed an interrupted SA probe action* identified how often operators were able to return back to their previous task. A system that is easy to remember is desirable in order to attain optimal operator behavior [90]. The percentage of completed interrupted SA probe actions mean (SD) are presented in Table 4.49. IA operators using the M_3 model were able to complete 100% of their interrupted actions compared to those using the M_2 model, while Collective operators using the M_2 model completed approximately 99% of their interrupted actions. The percentage of completed interrupted SA probe actions median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.27. Significant

differences existed between models for the IA operators for SA_O, while no differences existed for the Collective operators. Additional between visualizations Mann-Whitney-Wilcoxon tests identified a significant effect when using the M_3 model for SA_1 (n = 253, U = 55608, ρ = 0.03). Operators using the IA visualization completed more interrupted actions compared those using the Collective visualization. No correlations were found between the completed interrupted SA probe actions and SA probe accuracy.

Table 4.50: Investigate commands per decision mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	2.1 (3.23)	1.78 (1.62)
	Easy	2.06 (2.75)	1.53 (1.49)
	Hard	2.15 (3.79)	2.06 (1.72)
M_3	Overall	8.72 (3.82)	4.74 (2.2)
	Easy	8.09 (3.95)	4.23 (2.11)
	Hard	9.64 (3.44)	5.47 (2.12)

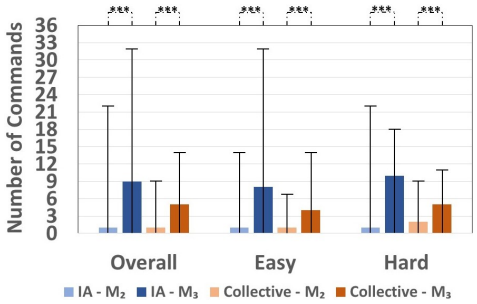


Figure 4.28: The number of investigate commands issued per decision median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The *investigate* command permitted increasing a collective's support for an operator specified target. Additional support for the same target was achieved by reissuing the investigate command repeatedly. The number of investigate commands issued per decision mean (SD) are presented in Table 4.50 [171]. Generally, operators using the M_2 model and Collective visualization issued fewer investigate commands. The number of investigate commands issued per decision median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.28. Significant differences were found between models for the number of investigate commands issued for both visualizations at all decision difficulties. Additional between visualizations

Mann-Whitney-Wilcoxon tests identified a moderate significant effect when using the M_2 model for overall decisions (n = 672, U = 63866, $\rho < 0.01$) and a highly significant effect for hard decisions (n = 298, U = 14066, $\rho < 0.001$). Highly significant effects between visualizations using the M_3 model were found for overall (U = 17990, $\rho < 0.001$), easy (n = 396, U = 6279.5, $\rho < 0.001$), and hard decisions (n = 276, U = 2331.5, $\rho < 0.001$).

Table 4.51: Abandon commands per decision mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	0.1 (0.54)	0.09 (0.29)
	Easy	0.05 (0.22)	0.06 (0.24)
	Hard	0.16 (0.79)	0.12 (0.34)
M_3	Overall	0.15 (0.43)	0.17 (0.42)
	Easy	0.15 (0.45)	0.16 (0.4)
	Hard	0.15 (0.4)	0.19 (0.45)

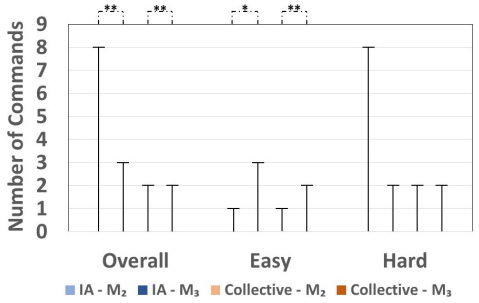


Figure 4.29: The number of abandon commands issued per decision median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The *abandon* command permitted decreasing a collective's support for a target and only needed to be issued once in order for the collective to ignore a specified target for the duration of a decision. The number of abandon commands issued per decision mean (SD) are shown in Table 4.51 [171]. Operators using the M_2 model in general issued fewer abandon commands compared to the M_3 model; however, IA operators using the M_3 model issued fewer abandon commands for hard decisions. The number of abandon commands issued per decision median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.29. Significant differences were found between models for the number of abandon commands issued per decision with both visualizations for overall and easy decisions. No significant effects

between visualizations were found. IA operators issued fewer abandon commands in general compared to those using the Collective visualization. Collective operators using the M_2 model issued fewer abandon commands for overall and hard decisions only.

Table 4.52: Decide commands per decision mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	0.38 (0.49)	0.52 (0.51)
	Easy	0.38 (0.49)	0.58 (0.51)
	Hard	0.39 (0.49)	0.44 (0.51)
M_3	Overall	0.99 (0.08)	1.03 (0.26)
	Easy	1 (0.07)	1.03 (0.24)
	Hard	0.99 (0.09)	1.04 (0.29)

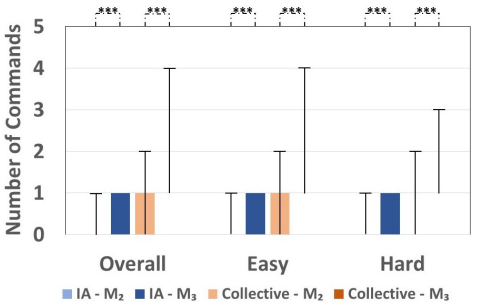


Figure 4.30: The number of decide commands issued per decision median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

A collective's entities stopped exploring alternative targets and moved to the operator selected target when the *decide* command was issued. A decide request required at least 30% of the collective support for the operator specified target. Collectives that reached 50% support for a target transitioned into the executing state and the operator was no longer able to influence the collective behavior. The number of decide commands issued per decision mean (SD) are presented in Table 4.52 [171]. Operators using the M_2 model with the IA visualization issued fewer decide commands compared to those using the M_3 model or the Collective visualization. The number of decide commands issued per decision median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.30. Significant differences were found between models for the number of decide commands issued per decision for both visualizations at all decision difficulties. Additional between visualizations

Mann-Whitney-Wilcoxon tests identified highly significant effects using the M_2 model for overall (n = 672, U = 63968, $\rho < 0.01$) and easy decisions (n = 374, U = 21014, $\rho < 0.001$). A moderately significant effect between visualizations when using the M_3 model was found for overall decisions (U = 57952, $\rho < 0.01$) and a significant effect existed for easy decisions (n = 377, U = 19997, $\rho = 0.05$).

Collective right-clicks and *target right-clicks* allowed the operator to access the respective information pop-up windows, which provided the number of individual entities in each particular state and the percentage of support each collective had for a respective target. The M_3 model in general had fewer collective and target right-clicks compared to the M_2 model, while the Collective visualization had fewer target right-clicks compared to the IA visualization. The statistical analyses were provided in Chapter 4.2.2.1.

Metrics showing how operators used the abandon command were assessed. Operators using the M_3 model and IA visualization abandoned the *highest value target* less frequently and had fewer *abandoned target information pop-up windows open*. The statistical analyses of both metrics were provided in Chapter 4.2.2.1. Instances may have occurred when the operator accidentally issued an undesired abandon command or repeatedly issued the abandon command, although targets were abandoned after a single command was issued. The percent of times *abandon commands exceeded abandoned targets* was examined and the mean (SD) are presented in Table 4.53 [177]. Operators using the M_2 model issued fewer repeated abandon commands compared to the M_3 model. The percent of times abandon commands exceeded abandoned targets median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.31. Significant differences were found between models for the percent of times abandon commands exceeded abandoned targets with both visualizations for overall and hard decisions. No significant effects between visualizations were found. IA oper-

ators had fewer repeated abandon commands in general compared to those using the Collective visualization. Collective operators using the M_3 model had fewer repeated abandon commands for overall and hard decisions.

Table 4.53: The percentage of times abandon commands exceeded abandoned targets per participant mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	1.18 (3.02)	2.68 (6.27)
	Easy	0.4 (1.55)	2.05 (5.06)
	Hard	1.35 (4)	3.08 (7.74)
M_3	Overall	6.88 (6.62)	6.54 (6.32)
	Easy	1.48 (4.39)	2.82 (5.73)
	Hard	13.26 (9.85)	10.91 (9.38)

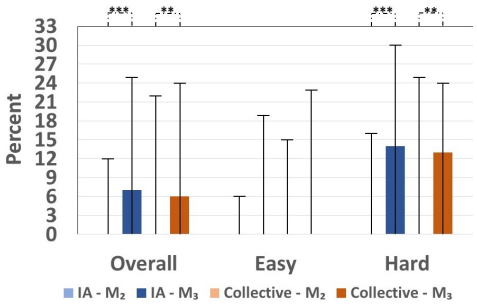


Figure 4.31: The percent of times abandon commands exceeded abandoned targets median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The *time difference (minutes) between the commit state and issued decide command* assessed the operator's ability to predict the collective's future state transition from the committed state (30% support for a target) to executing (50% support for a target). The time difference mean (SD) are shown in Table 4.54 [177]. Operators using the M_3 model issued decide commands faster than the M_2 model. The time difference between commit state and issued decide command median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.32. Significant differences existed between models for the time difference between the commit state and decide command for both visualizations at all decision difficulties. Collective operators in general had smaller time differences between the committed state and issued decide commands compared to those using the IA visualization; however, no significant effects between visualizations were found. IA operators using the M_2 model had smaller

time differences between the commit state and decide command for hard decisions.

Table 4.54: The time difference (minutes) between commit state and issued decide command per participant mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	0.68 (0.27)	0.65 (0.15)
	Easy	0.7 (0.47)	0.56 (0.14)
	Hard	0.72 (0.21)	0.78 (0.3)
M_3	Overall	0.6 (0.3)	0.57 (0.2)
	Easy	0.57 (0.53)	0.52 (0.32)
	Hard	0.62 (0.22)	0.62 (0.18)

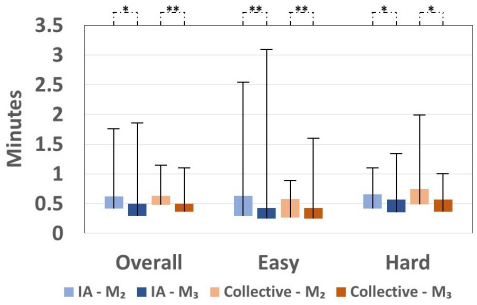


Figure 4.32: The time difference between commit state and issued decide command median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

Further analysis of how operators used the collective and target information pop-up windows was conducted. The average number of times target information pop-up windows were opened per target per decision identified the *average frequency at which the information pop-up windows were accessed*. The average frequency of an accessed target information pop-up window per target per decision mean (SD) are shown in Table 4.55. Operators using the M_3 model in general accessed target information pop-up windows less frequently compared to the M_2 model. Target information pop-up windows were accessed less frequently for operators from both evaluations using the M_2 model for easy decisions. The average frequency of an accessed target information pop-up window median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.33. IA operators had significantly different average frequencies of accessed target information pop-up windows between models for hard decisions, while the Collective operators had no significant differences between models. Additional between visualizations Mann-Whitney-Wilcoxon tests identified a sig-

nificant effect when using the M_2 model for overall decisions (n = 619, U = 42857, ρ = 0.02) and a moderate significant effect for hard decisions (n = 282, U = 7908.5, $\rho <$ 0.01). Operators using the Collective visualization accessed target information pop-up windows less frequently compared to the IA visualization.

Table 4.55: Average frequency of accessed target information pop-up window per target per decision mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	1.93 (1.17)	1.67 (0.94)
	Easy	1.7 (0.98)	1.57 (0.82)
	Hard	2.23 (1.33)	1.79 (1.05)
M_3	Overall	1.8 (1.33)	1.67 (0.91)
	Easy	1.83 (1.48)	1.62 (0.88)
	Hard	1.77 (1.1)	1.73 (0.95)

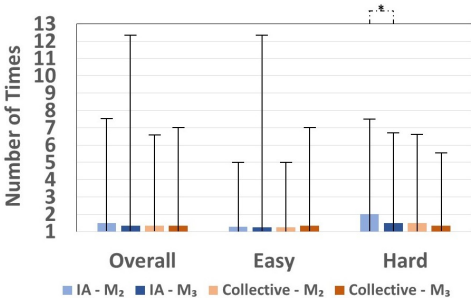


Figure 4.33: Average frequency of accessed target information pop-up window per target median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

Operators using the target information pop-up windows may have accessed them frequently for short time periods, or left them open for longer time periods. The average percentage of *time a target information pop-up window was open per target* relative to the decision time mean (SD) are presented in Table 4.56. IA operators using the M_2 model left target information pop-up windows open for shorter time periods. The average time target information windows were opened median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are presented in Figure 4.34. Significant differences were found between models for the average time target information pop-up windows were open for both visualizations at all decision difficulties; however, no significant effects between visualizations were found.

Table 4.56: Average time target information windows opened per target per decision (%) mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	24.18 (26.65)	28.38 (28.61)
	Easy	27.53 (28.76)	30.48 (29.12)
	Hard	19.87 (23.05)	26.05 (27.95)
M_3	Overall	34.93 (25.29)	36.58 (29.41)
	Easy	37.56 (27.01)	37.63 (30.98)
	Hard	31.12 (22.12)	35.09 (27.07)

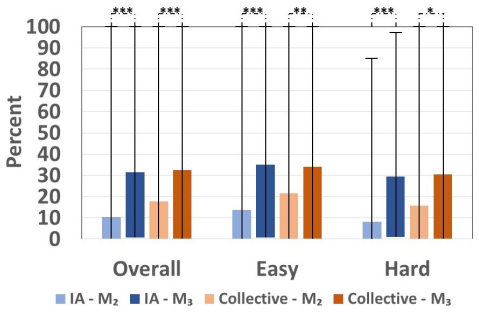


Figure 4.34: Average time target information windows opened per target median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

Operators may have accessed particular target information pop-up windows, such as the decision target, more frequently for longer time periods. The average percentage of *time the decision target information pop-up window was open* relative to the decision time mean (SD) are shown in Table 4.57. Operators using the M_2 model left the decision target information pop-up window open for shorter periods of time compared to the M_3 model. The time the decision target information window is open median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.35. Significant differences were found between models for the time the decision target information window was open for both visualizations at all decision difficulties. Additional between visualizations Mann-Whitney-Wilcoxon tests identified a highly significant effect using the M_2 model for overall decisions (n = 672, U = 65102, $\rho < 0.001$), as well as significant effects for easy (n = 374, U = 20114, $\rho = 0.01$), and hard decisions (n = 298, U = 12832, $\rho = 0.02$). A moderately significant effect between visualizations using the M_3 model was found for overall decisions (U = 48749, $\rho < 0.01$), with significant effects for easy (n = 396, U = 17095, $\rho = 0.03$), and hard decisions (n = 276,

U = 8157, ρ = 0.04). IA operators using the M_2 model left the decision target information pop-up window open for shorter periods of time compared to those using the M_3 model, while the Collective operators had shorter time periods using the M_3 model.

Table 4.57: The time decision target information window open per decision (%) mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	21.64 (28.25)	30.55 (32.6)
	Easy	23.69 (30.7)	32.51 (34.43)
	Hard	18.84 (24.33)	28.27 (30.31)
M_3	Overall	50.56 (29.1)	43.94 (31.69)
	Easy	50.71 (29.31)	44.12 (33.33)
	Hard	50.34 (28.91)	43.67 (29.31)

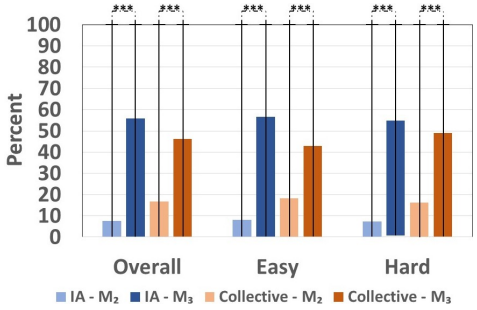


Figure 4.35: The time decision target information window open median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The average percentage of *time the decision collective information pop-up window was open* relative to the decision time mean (SD) are shown in Table 4.58. The time the decision collective information window was open was only assessed for the IA evaluation, because the Collective evaluation did not record which particular collective pop-up window was opened or closed. IA operators using the M_3 model left the decision collective information pop-up window open for shorter periods of time compared to the M_2 model. The time the decision collective information window is open median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.36. IA operators had significantly different times for hard decisions.

The post-trial questionnaire assessed the *perceived effectiveness of each request type* (investigate, abandon, and decide), not effective (1) to very effective (7). The post-trial effectiveness subjective ranking mean (SD) are presented in Table 4.59 [29]. The in-

Table 4.58: The time decision collective information window open per decision (%) mean (SD) by decision difficulty (Dec Diff).

	Dec Diff	IA
M_2	Overall	21.37 (35.24)
	Easy	20.16 (34.79)
	Hard	23.03 (35.91))
M_3	Overall	19.84 (35.28)
	Easy	19.74 (34.79)
	Hard	20 (36.12)

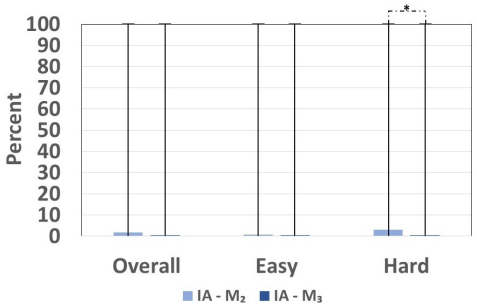


Figure 4.36: The time decision collective information window open median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

vestigate, abandon, and decide rankings were generally ranked higher for operators using the M_3 model when compared to those using the M_2 model. Collective operators using the M_2 model ranked abandon effectiveness higher. The post-trial effectiveness median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.37. Significant differences between models were found in IA operator rankings for the decide command and for Collective operator rankings for both the abandon and decide commands. Additional between visualizations Mann-Whitney-Wilcoxon tests identified a moderate significant effect for the abandon effectiveness when using the M_2 model (n = 56, U = 554.5, $\rho < 0.01$). IA operators using the M_3 model ranked investigate, abandon, and decide effectiveness higher compared to those using the Collective visualization, while Collective operators ranked abandon effectiveness higher when using the M_2 model.

The post-experiment questionnaire assessed the collective's *responsiveness* to requests, the participants' *ability* to choose the highest valued target, and their *understanding* of the collective behavior. IA operators who used the M_2 model had the best collective

Table 4.59: Post-trial command effectiveness ranking mean (SD) (1-low, 7-high).

	Metric	IA	Collective
M_2	Investigate	4.68 (1.56)	4.75 (1.53)
	Abandon	4.82 (1.96)	6.18 (1.42)
	Decide	5.29 (1.7)	5.57 (1.99)
M_3	Investigate	5.46 (1.4)	5.18 (1.68)
	Abandon	5.29 (1.84)	5.29 (1.76)
	Decide	6.79 (0.5)	6.54 (0.92)

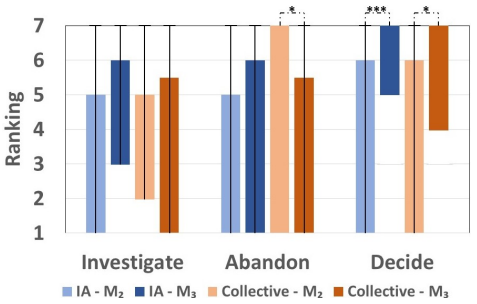


Figure 4.37: Post-trial command effectiveness ranking median (min/max) and Mann-Whitney-Wilcoxin test between models.

responsiveness, operator ability, and understanding versus the M_3 model. Collective operators ranked the collective's responsiveness highest using the M_3 model, while operator ability and understanding were highest using the M_2 model. Details regarding the statistical tests were provided in the Metrics and Results Chapter 4.2.1.1.

Table 4.60: A synopsis of R_7's hypotheses associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question.

Variable	Sub-Variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Global Clutter Percentage	SA_O	H_{14} -AT		H_{14} -AT					H_{14} -B
	SA_1	H_{14} -AT		H_{14} -AT				H_{14} -B	
	SA_2	H_{14} -AT	H_{14} $-D$	H_{14} -B,W					
	SA_3								H_{14} -B

Variable	Sub-Variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Euclidean Distance Between SA Probe Interests and Clicks	SA_O	H_{15} -AT		H_{15} $-AT$					H_{15} -W,D
	SA_1			H_{15} -B,W		H_{15} $-B$			
	SA_2	H_{15} -AT	H_{15} -B						
Middle of Action During SA Probe	SA_O	H_{15}	H_{15}	H_{15}	H_{15}				
	SA_1	H_{15}	H_{15}	H_{15}	H_{15}	H_{15}			
	SA_2	H_{15}	H_{15}	H_{15}	H_{15}		H_{15}		
	SA_3			H_{15}	H_{15}		H_{15}	H_{15}	
Completed Interrupted SA Probe Action	SA_O	H_{14}, H_{15}							
	SA_1				H_{14}, H_{15}				
Investigate Commands	Overall	H_{15}	$\mathbf{H_{15}}$	$\mathbf{H_{15}}$	H_{15}				
	Easy	H_{15}	$\mathbf{H_{15}}$		H_{15}				
	Hard	H_{15}	$\mathbf{H_{15}}$	$\mathbf{H_{15}}$	H_{15}				
Abandon Commands	Overall	H_{15}	H_{15}						
	Easy	H_{15}	H_{15}						
Decide Commands	Overall	H_{14}, H_{15}	H_{14}, H_{15}	H_{14}, H_{15}	H_{14}, H_{15}				
	Easy	H_{14}, H_{15}	H_{14}, H_{15}	H_{14}, H_{15}	H_{14}, H_{15}				
	Hard	H_{14}, H_{15}	H_{14}, H_{15}						
Collective Right-Clicks	Overall	H_{15}							
	Hard	H_{15}							
Target Right-Clicks per Decision	Easy	H_{15}							
Highest Value Target Abandoned	Overall		H_{14}						
	Easy	H_{14}							

Variable	Sub-variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Abandoned Target Info. Window Open	Overall				H_{14}				
	Easy				H_{14}				
Abandon Requests Exceeded Abandon Targets	Overall	H_{14}	H_{14}						
	Hard	H_{14}	H_{14}						
Time Between Commit State Issued Decide Command	Overall	H_{14}	H_{14}						
	Easy	H_{14}	H_{14}						
	Hard	H_{14}	H_{14}						
Frequency of Accessed Target Info. Windows	Overall			H_{14}, H_{15}					
	Hard	H_{14}, H_{15}		H_{14}, H_{15}					
Time Target Info. Windows Open	Overall	H_{14}	H_{14}						
	Easy	H_{14}	H_{14}						
	Hard	H_{14}	H_{14}						
Time Decision Collective Info. Window Open	Hard	H_{14}							
Time Decision Target Info. Window Open	Overall	H_{14}	H_{14}	H_{14}	H_{14}				
	Easy	H_{14}	H_{14}	H_{14}	H_{14}				
	Hard	H_{14}	H_{14}	H_{14}	H_{14}				
Post-Trial	Abandon		H_{14}	H_{14}					
	Decide	H_{14}	H_{14}						
Post-Experiment	Respon.	H_{14}	H_{14}						

A summary of R_7's results by the hypotheses, with significant results identified, is provided in Table 4.60. This summary table is intended to facilitate the discussion.

4.2.3.2 Discussion

The analysis of which model and visualization promoted better *usability* suggests that the IA visualization promoted transparency more *effectively* than the Collective visualization, while both models had their respective advantages and disadvantages. Operators using the M_2 model had less global clutter, due to target *information* pop-up windows being open for less time, smaller Euclidean distances between the interest of a *SA* probe question and their current interaction, were able to complete interrupted actions after answering a *SA* probe question, and issued fewer abandon and decide commands. H_{14}, which hypothesized that the M_2 model and Collective visualization will promote better *usability* by being more *predictable* and *explainable*, was not supported by the M_2 model results. Operators from both evaluations using the M_2 model abandoned the highest value target more frequently, which may have occurred due to mis*understanding* or poor *SA*. IA operators using the M_2 model were not as *timely* (i.e., faster) at *predicting* when a collective was committed to a target and had the decision collective *information* pop-up window open for a longer duration of time (i.e., lower *explainability*) compared to when using the M_3 model. The Collective evaluation did not record which collectives were right-clicked on, which impeded the ability to associate right-clicks to a collective per decision; however, a similar reliance of having the decision collective *information* pop-up window visible, like the IA operators, may have occurred considering how the Collective operators used the target *information* pop-up windows. Further evaluations are needed to validate Collective operator *usability* behavior.

The Collective visualization enabled operators to complete actions prior to a SA probe question and aided operators to issue decide commands shortly after a collective was committed to a target. H_{14} was not supported by the Collective visualization find-

ings, since more of the highest value targets were abandoned. The continuous display of collective and target *information* pop-up windows promoted higher *SA performance* for the Collective operators when using both models. The reliance of the supplementary *information* provided in the pop-up windows suggests that the *information* was more *explainable* and *reliable* than the information provided on the collective icons. Incorporating the numerical percentage of support from the respective Collectives on a target icon or identifying the most favored target on a collective hub may help reduce the reliance of the *information* pop-up windows and simultaneously improve *SA* by mitigating potential *observability* issues if the operator must interact with more collectives.

IA operators using the M_3 model and Collective operators using the M_2 model were able to complete actions that were interrupted by a SA probe question 99% of the time. The *memorability* of both models and visualizations enabled operators to return to a previous task after answering the *SA* probe question, because of the required operator engagement (M_3 model) and established expectations of collective behaviors (M_2 model). The *predictability* of the M_3 model and Collective visualization *justified* issuing decide commands shortly after collectives were in a committed state; however, this finding may be biased for the M_3 model, because of the required operator influence to achieve the decision-making task. The same bias can attribute to the command *effectiveness* rankings, which were higher for the M_3 model. The IA operators' ability to identify objects on the visualization may have been impeded by displaying all of the individual collective entities, collective and target icons, and collective and target *information* pop-up windows when the *SA* probe question inquired about an object further away from the center of the operator's current attentional focus. Asking SA probe questions about objects at various distances from the operator's current focal point is necessary in order to *understand* how clutter, or moving individual collective entities, may affect

the operator's ability to identify the object of interest and impact *SA performance*.

H_{15}, which hypothesized that operators using the M_2 model and Collective visualization will require fewer interactions, was not supported. The M_2 model enabled fewer commands compared to the M_3 model, as expected. The IA visualization enabled fewer abandon and decide commands. Collective operators using the M_2 model had better decision-making *performance* when more investigate commands were issued. Issuing more investigate commands for high-value targets located further away from the collective hub may suggest that the interaction delay embedded in the M_2 model, which was designed to reduce the impacts of environmental bias and improve the success of choosing the ground truth best targets, may have not accommodated operators' expectations if lower valued targets were being favored solely because they were closer to the hub. Collective operators who issued more commands may have wanted *control* and *directability* over the decision-making, which may have occurred due to lower trust, or mis*understanding* collective behavior. Further investigations are needed to determine if and how trust may influence operators. Operators implemented different strategies to fulfill the decision-making task; however, the most successful strategy promoted more consensus decision-making (i.e., investigate commands), as opposed to prohibiting exploration of targets (i.e., abandon commands). *Understanding* how operators used commands is necessary to promote *effective* interactions and produce desired human-collective *performance*.

The transparency embedded in the M_2 model and Collective visualization combination did not support the best overall system *usability*. The IA visualization promoted less clutter, by alleviating the dependence of the collective and target *information* pop-up windows, and promoted fewer interactions. Modifications to both the M_2 model and Collective visualization must be made in order to mitigate the highest value target

being abandoned more frequently, as well as reduce the reliance on the *information* windows. The assumption that fewer interactions are optimal may not be accurate for all decision difficulties, such as hard decisions. *Understanding* strategies and *justifications* for more interactions is necessary in order to promote transparency that aids operators during particular situations and results in higher human-collective *performance*.

4.2.4 R_8: System Design Element Influence on Team Performance

Assessing *which model and visualization promoted better human-collective performance*, R_8, is necessary to determine whether the human-collective system transparency aided the task. The objective dependent variables were (1) decision time, (2) selection success rate, and (3) SA probe accuracy.

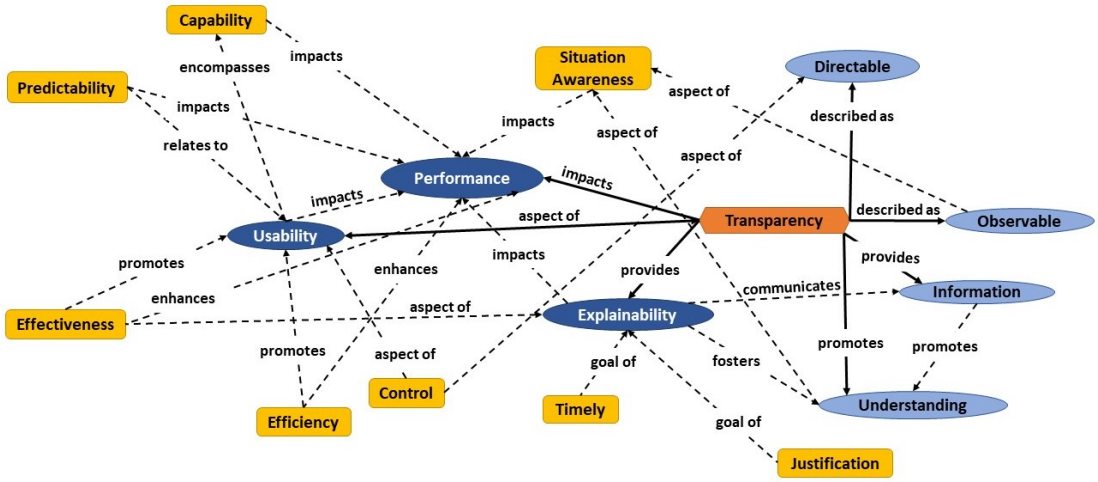


Figure 4.38: R_8 concept map of the assessed direct and indirect transparency factors.

Objective metrics were included to support the correlation analyses. The specific direct and indirect transparency factors related to R_8 are identified in Figure 4.38. The

relationship between the variables and the corresponding hypotheses, as well as the direct and indirect transparency factors, are identified in Table 4.61.

Performance of the human-collective team can be used to assess the effects of the model and visualization transparency on the team's ability to fulfill tasks. It was hypothesized (H_{16}) that the human-collective performance, effectiveness, efficiency, and timing will be better using the M_2 model with the Collective visualization.

Table 4.61: Interaction of system design elements influence on human-collective performance objective (obj) and subjective (subj) variables (vars), relationship to hypothesis H_{16}, as well as the associated direct and indirect transparency factors, are presented in Figure 2.2.

	Transparency Factors														
	Direct							**Indirect**							
Obj Vars	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Capability**	**Control**	**Effectiveness**	**Efficiency**	**Justification**	**Predictability**	**SA**	**Timing**
Dec Time					✓					✓	✓				✓
Selection Success Rate					✓					✓					
SA Probe Accuracy				✓	✓	✓				✓			✓	✓	
Collective Observation		✓	✓				✓			✓		✓			
Target Observation		✓				✓	✓								
Collective Right-Clicks		✓	✓				✓			✓		✓			
Target Right-Clicks		✓	✓				✓			✓		✓			
Investigate Commands					✓		✓		✓	✓					

	Transparency Factors														
	Direct							Indirect							
Obj Vars	**Directable**	**Explainability**	**Information**	**Observable**	**Performance**	**Understanding**	**Usability**	**Capability**	**Control**	**Effectiveness**	**Efficiency**	**Justification**	**Predictability**	**SA**	**Timing**
Abandon Commands					✓		✓		✓	✓					
Decide Commands					✓		✓		✓	✓			✓		
Time Dec Target Info Wind Open							✓			✓					✓
Time Dec Coll Info Wind Open							✓			✓					✓
Mental Rotation Assessment					✓			✓							
Working Memory Capacity					✓			✓							
Subj Vars															
Weekly Hours on a Desktop or Laptop					✓			✓							
Post-Trial Performance Understanding		✓			✓	✓		✓							

4.2.4.1 Metrics and Results

The length of time the human-collective team reached a decision, *decision time* (minutes), was examined. The decision time mean (SD) are shown in Table 4.62 [57, 171, 177].

Collective operators using the M_2 model had the fastest decision times. The decision time median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.39. Significant differences were found between models for both visualizations at all decision difficulties. Additional between visualizations Mann-Whitney-Wilcoxon tests identified significant effects when using the M_2 model for overall (n = 672, U = 50921, ρ = 0.03), easy (n = 375, U = 15452, ρ = 0.04), and hard decisions (n = 297, U = 9521, ρ = 0.04). A significant effect between visualizations using the M_3 model was also found for easy decisions (n = 396, U = 17376, ρ = 0.05).

Table 4.62: Decision time (minutes) mean (SD) per decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	4.32 (1.83)	3.97 (1.37)
	Easy	3.77 (1.63)	3.37 (1.23)
	Hard	5.09 (1.82)	4.67 (1.2)
M_3	Overall	5.67 (2.86)	5.32 (2.22)
	Easy	5.22 (3.06)	4.67 (1.96)
	Hard	6.32 (2.42)	6.24 (2.24)

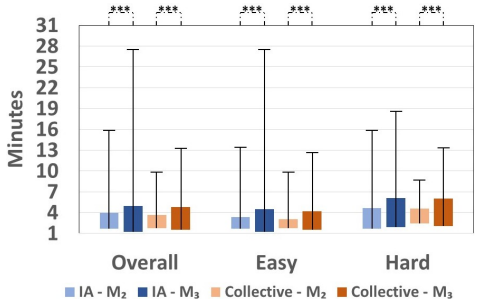


Figure 4.39: Decision time median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The *selection success rate* was the number of correct decisions (the collective moved to the highest valued target) relative to the total number of decisions. Selection success rate mean (SD) per decision are shown in Table 4.63 [57, 171, 177]. Operators using the M_3 model with the Collective visualization in general had higher selection success rates, while IA operators using the M_2 model had higher selection success rates for hard decisions. The selection success rate median, min, max, and the Mann-Whitney-Wilcoxon significant effects between models are shown in Figure 4.40. Collective operators had significant differences in selection success rate between models for overall

decisions, while no significant differences between models were found for IA operators. Additional between visualizations Mann-Whitney-Wilcoxon tests identified highly significant effects when using the M_2 model for overall (n = 672, U = 64008, $\rho < 0.001$) and easy decisions (n = 375, U = 19845, $\rho < 0.001$), as well as a moderate significant effect for hard decisions (n = 297, U = 12761, $\rho < 0.01$). Highly significant effects between visualizations using the M_3 model for overall (U = 66360, $\rho < 0.001$, easy (n = 396, U = 21662, $\rho < 0.001$), and hard decisions (n = 276, U = 12178, $\rho < 0.01$) existed. The Spearman correlation analysis revealed a moderate correlation between decision time and selection success rate using the M_2 model with the IA visualization for easy decisions (r = -0.42, $\rho < 0.001$) and a weak correlation for overall decisions (r = -0.27, $\rho < 0.001$). Weak correlations existed when using the M_2 model with the Collective visualization for overall (r = -0.11, $\rho = 0.05$), easy (r = -0.18, $\rho = 0.02$), and hard decisions (r = 0.18, $\rho = 0.03$). A weak correlation was found for hard problems using the M_3 model with the IA (r = 0.32, $\rho < 0.001$) and Collective visualizations (r = 0.25, $\rho < 0.01$).

Table 4.63: Selection success rate (%) mean (SD) per decision difficulty (Dec Diff).

	Dec Diff	IA	Collective
M_2	Overall	75 (43.37)	88.39 (32.08)
	Easy	81.44 (38.98)	94.44 (22.97)
	Hard	66.2 (47.47)	81.41 (39.03)
M_3	Overall	75.3 (43.19)	92.86 (25.79)
	Easy	85.43 (35.37)	95.94 (19.79)
	Hard	60.58 (49.05)	88.49 (32.03)

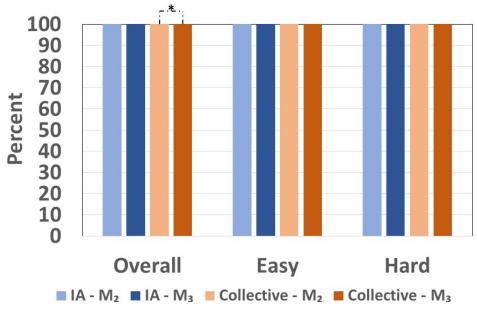


Figure 4.40: Selection success rate median (min/max) and Mann-Whitney-Wilcoxin test by decision difficulty between models.

The IA and Collective operators' *SA probe accuracy* when using the M_2 model was higher for SA_3, while the IA operators had higher SA_2 and the Collective operators had

higher SA_O. Collective operators using either model had higher SA probe accuracy. The detailed statistical analyses were provided in Chapter 4.2.1.1.

Additional Spearman correlation analyses analyzed if any correlations existed between selection success rate and some objective metrics, including collective and target observations and right-clicks, investigate, abandon, and decide commands, as well as the time a decision collective and target information pop-up window was open. A weak correlation existed for collective observations using the Collective visualization with the M_2 model for overall decisions (r = -0.12, ρ = 0.03). Weak correlations were found for target observations when using the Collective visualization with the M_3 model for overall (r = 0.14, ρ = 0.01) and hard decisions (r = 0.16, ρ = 0.05). Weak correlation were found for the number of target right-clicks using the IA visualization with the M_2 model for overall decisions (r = -0.13, ρ = 0.02), and with the M_3 model for overall (r = 0.1, ρ = 0.05) and hard decisions (r = 0.18, ρ = 0.03), as well as when using the Collective visualization with the M_2 model for hard decisions (r = 0.17, ρ = 0.04). Weak correlations were found for the number of investigate commands when using the Collective visualization with the M_2 model for hard decisions (r = 0.2, ρ = 0.01), as well as when using the IA visualization with M_3 model for easy (r = -0.16, ρ = 0.02) and hard decisions (r = 0.24, $\rho < 0.01$). Weak correlations were found for the number of abandon commands when using the IA visualization with the M_2 model for easy decisions (r = -0.19, $\rho <$ 0.01), and with the M_3 model for hard decisions (r = 0.2, ρ = 0.02). A weak correlation existed for the number of decide commands using the Collective visualization with the M_3 model for overall decisions (r = 0.11, ρ = 0.05). Weak correlations were found for the time a decision target information pop-up window was open when using the Collective visualization with the M_2 model for overall (r = 0.11, ρ = 0.04) and hard decisions (r = 0.16, ρ = 0.04). No significant effects were found for collective right-clicks and the time

a decision collective information pop-up window was open.

Spearman correlation analyses were also conducted to identify correlations between selection success rate and some subjective metrics, including the weekly hours on a desktop or laptop, the mental rotations assessment, and working memory capacity. Weak correlations were found for the weekly hours participants' used a desktop or laptop for easy decisions when using the IA visualization with the M_2 model (r = 0.16, ρ = 0.02) and with the M_3 model (r = -0.15, ρ = 0.04), as well as when using the Collective visualization with the M_3 model for hard decisions (r = 0.17, ρ = 0.05). A weak correlation was found for the mental rotations assessment using the IA visualization with the M_3 model for hard decisions (r = 0.18, ρ = 0.04). Weak correlations were found for working memory capacity and easy decisions when using the IA visualization with the M_2 model (r = -0.17, ρ = 0.02), and with the M_3 model (r = -0.15, ρ = 0.04).

The *post-trial performance and understanding* questionnaire assessed the operators' understanding of the collectives' behavior and their ability to choose the best target. The Collective operators ranked performance and understanding higher when using the M_3 model. The statistical analysis details were provided in Chapter 4.2.2.1.

Table 4.64: A synopsis of R_8's hypotheses associated with significant results. The SA probe timings are all timings (AT), 15 seconds Before asking (B), While being asked (W), and During response (D) to a SA probe question.

Variable	**Sub-variable**	**Within Model**		**Between Visualization**		**Correlation**			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Decision Time	Overall	H_{16}	$\boldsymbol{H_{16}}$	$\boldsymbol{H_{16}}$		H_{16}		$\boldsymbol{H_{16}}$	
	Easy	H_{16}	$\boldsymbol{H_{16}}$	$\boldsymbol{H_{16}}$	H_{16}	H_{16}		$\boldsymbol{H_{16}}$	
	Hard	H_{16}	$\boldsymbol{H_{16}}$	$\boldsymbol{H_{16}}$			H_{16}	$\boldsymbol{H_{16}}$	H_{16}

Variable	Sub-variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Selection Success Rate	Overall		H_{16}	H_{16}	H_{16}				
	Easy			H_{16}	H_{16}				
	Hard			H_{16}	H_{16}				
SA Probe Accuracy	SA_O			$\mathbf{H_{16}}$	H_{16}				
	SA_1	H_{16}		$\mathbf{H_{16}}$	H_{16}				
	SA_2			$\mathbf{H_{16}}$	H_{16}				
	SA_3			$\mathbf{H_{16}}$	H_{16}				
Collective Observations	Overall	H_{16}		H_{16}	H_{16}			$\mathbf{H_{16}}$	
	Easy	H_{16}	H_{16}	H_{16}	H_{16}				
	Hard	H_{16}			H_{16}				
Target Observations	Overall	H_{16}	$\mathbf{H_{16}}$	$\mathbf{H_{16}}$	H_{16}				H_{16}
	Easy		$\mathbf{H_{16}}$	$\mathbf{H_{16}}$	H_{16}				
	Hard		$\mathbf{H_{16}}$	$\mathbf{H_{16}}$	H_{16}				H_{16}
Collective Right-Clicks	Overall	H_{16}							
	Hard	H_{16}							
Target Right-Clicks per Decision	Overall					H_{16}	H_{16}		
	Easy	H_{16}							
	Hard						H_{16}	$\mathbf{H_{16}}$	
Investigate Commands	Overall	H_{16}	$\mathbf{H_{16}}$	$\mathbf{H_{16}}$	H_{16}				
	Easy	H_{16}	$\mathbf{H_{16}}$		H_{16}		H_{16}		
	Hard	H_{16}	$\mathbf{H_{16}}$	$\mathbf{H_{16}}$	H_{16}		H_{16}	$\mathbf{H_{16}}$	
Abandon Commands	Overall	H_{16}	H_{16}						
	Easy	H_{16}	H_{16}			H_{16}			
	Hard						H_{16}		
Decide Commands	Overall	H_{16}	H_{16}	H_{16}	H_{16}				H_{16}
	Easy	H_{16}	H_{16}	H_{16}	H_{16}				
	Hard	H_{16}	H_{16}						
Time Decision Collective Info. Window Open	Hard	H_{16}							
Time Decision Target Info. Window Open	Overall	H_{16}	H_{16}	H_{16}	H_{16}			$\mathbf{H_{16}}$	
	Easy	H_{16}	H_{16}	H_{16}	H_{16}				
	Hard	H_{16}	H_{16}	H_{16}	H_{16}			$\mathbf{H_{16}}$	

Variable	Sub-variable	Within Model		Between Visualization		Correlation			
		IA	Coll.	M_2	M_3	IA		Coll.	
						M_2	M_3	M_2	M_3
Mental Rotations Assessment	SA_O					H_{16}	H_{16}		
	SA_1					H_{16}	H_{16}		
	SA_2						H_{16}		
	Hard						H_{16}		
Working Memory Capacity	SA_O					H_{16}	H_{16}		
	SA_1					H_{16}			
	SA_3					H_{16}			
	Easy					H_{16}	H_{16}		
Weekly Hours on a Desktop or Laptop	SA_O					H_{16}			
	SA_1					H_{16}			
	SA_2							H_{16}	
	Easy					H_{16}	H_{16}		
	Hard								H_{16}

A summary of R_8's results that show the hypotheses with associated significant results is shown in Table 4.64. This summary table is intended to facilitate the discussion.

4.2.4.2 Discussion

The analysis suggests that the Collective visualization promoted better human-collective *performance*; however, the models had their respective advantages and disadvantages. The M_2 model promoted faster decision *times*, while the Collective visualization promoted faster decision *times*, higher selection success rates, and higher subjective *performance*. *SA performance* varied across the models and visualizations. H_{16}, which hypothesized that the human-collective *performance*, *effectiveness*, *efficiency*, and *timing* will be better using the M_2 model with the Collective visualization, was partially supported. Collective operators using the M_2 model had faster decision *times*; however, the M_3

model enabled higher selection success rates. Embedding transparency into the M_2 model requires (1) balancing *control* between the operator and the collectives so that the operators can positively contribute and *direct* decision-making, (2) promoting positive human-collective interactions so that the operator's and the collective's strengths are maximized, and (3) alleviating the operator's workload.

Understanding *usability* and what interactions were used by operators to *justify* actions that contributed to performance are necessary in order to identify the most *effective* and *efficient* strategies. Operators using the M_2 model issued fewer commands to the collectives, which was desired in order to maximize the usage of the collectives' consensus decision-making process; however, more particular interactions, such as investigate commands, resulted in higher selection success rate performance. Requiring operators to influence the task ensured better *performance*, because the operators were in-the-loop, versus operators who were supervising the collective behaviors and potentially correcting actions towards task success. Further analysis is required to determine how to improve target selection when using the M_2 model. Improvements during training may help emphasize the necessity of selecting the highest-value targets.

Realistic human-collective scenarios will require high *performance* with short decision *times*, especially in uncertain and dynamic environments. The design of an *effective* human-collective system must enable the human-collective team to fulfill primary objectives, without hindering other metrics, such as decision *time* and accuracy. Devoting more *time* to ensure high task *performance* is a common trade-off [179]. Expedited decisions may have occurred if higher valued targets were more *observable* further away from other objects (less clutter), making them more salient, or if impatient operators *predicted* future collective behaviors and influenced collectives more to make faster decisions. Using target outlines, collective and target *information* pop-up windows, and

issuing investigate commands were necessary to fulfill the primary task and can be used to ensure an *explainable* and *usable* system. The M_2 model with the Collective visualization enabled operators with different spatial *capabilities* to perform relatively the same, unlike IA operators, specifically those with lower Working Memory Capacity and more weekly desktop or laptop exposure, who had higher selection success rates.

The transparency embedded in the Collective visualization with the M_2 model promoted the fastest decision times; however, modifications are needed in order to improve the other human-collective *performance* metrics. *Understanding* what interactions contributed to higher *performance* is necessary to determine what operator strategies are most *effective* and *efficient*. The M_2 model subjective *performance* rankings may have had a consistent negative bias due to learning effects, since this model was always presented before using the M_3 model. Improving the transparency embedded in the Collective visualization to promote better *SA performance* must be considered. *Understanding* what IA visualization aspects, such as streamlines between collectives and targets, promoted better *SA performance* can be emulated in the more abstract Collective visualization.

4.2.5 Model with Visualization Analysis Discussion

The first research objective was to expand on the existing transparency literature by assessing how different models and visualizations influenced human-collective behavior. The analysis assessed *understanding* how the transparency embedded in the models and visualizations influenced operators with individual differences (i.e., *capabilities*), operator comprehension (i.e, *capability* of *understanding*), system design element *usability* (i.e., model and visualization usability), and human-collective *performance*. The second research objective was to determine whether using the best model with the best

visualization, derived from two previous analyses [57, 177], provided the best transparency. Previous results indicated that the M_2 model enabled faster decisions and relied less on operator influence [57], while the Collective visualization provided better transparency [177], because operators with different individual capabilities *performed* similarly for both tasks, and the human-collective team *performed* better. The M_2 model, independently, did not enable operators with individual differences to perform similarly; however, it did promote fewer interactions and less clutter, which enabled operators to complete interrupted actions, promoted faster decision times, and higher *SA performance*. The Collective visualization independently enabled operators with different individual differences to perform similarly, promoted higher *understanding* and *SA*, enabled operators to complete interrupted actions and issue decide commands shortly after a collective was committed to a target, promoted faster decision times, higher selection success rates, and higher subjective *performance*. Together the M_2 model with the Collective visualization promoted lower overall *workload*, required less physical demand, had fewer investigate commands and target observations (i.e., extra clicks), while enabling the fastest decision *time*. The different outcomes between the findings in this evaluation versus those from Cody *et al.* [57] and Roundtree *et al.* [177] suggest that transparency cannot be quantified by using the best system design elements, but instead must be quantified by considering how the transparency of the different system design elements interact with one another along with the implications of how that system transparency influences human-collective interactions and *performance*.

Fewer operator interactions was a desired behavior in order to minimize negative influence on collective behaviors and reduce the reliance on supplementary *information*; however, operator influence was anticipated to aid the decision-making process and *time* to complete decisions. This analysis identified positive and negative interactions

associated with both models and visualizations. Collective operators relied on visible target *information* windows more than 25% of the decision time, resulting in more global clutter. Clutter, from a system design perspective, can hinder *effective* task performance. Collective operators with more clutter were able to answer more *SA* probe questions correctly and had higher selection success rates. The dependence on visible collective and target *information* pop-up windows may have been influenced by the type of *SA* probe questions asked and the visualization not being observable without the supplementary *information*. Sixteen of twenty-four *SA* probe questions depended on numerical values of collective state and target support *information* provided in the collective and target *information* pop-up windows. Collective state *information* was provided via the different color individual collective entities on the IA visualization and the opacity of the Collective visualization's hub quadrants, while color and opacity were used to indicate the highest supporting collective on the target icon. The use of opacity may have been in*effective* and less salient; however, using different colors to indicate state *information* may be a possible design modification to the Collective visualization. Experimental design modifications can also be implemented in order to ensure a more even distribution of *SA* probe questions that rely on other *information*, such as the icons, system messages, or collective assignments versus *information* pop-up windows.

The use of target *information* pop-up windows aided Collective operators to abandon targets more than 25% of the *time*. Operators who used the target *information* pop-up windows to *justify* that a target was abandoned by a collective, may have been confused if the reported target support was not equal to zero. Additional abandon commands may have been issued in an attempt to reduce the collective support to zero. IA operators may have experienced a similar confusion if they observed individual collective entities still travelling to an abandoned target. Implementing design changes, such

as showing zero support when an abandon request has been committed, or not displaying lost entities after a specific period of *time*, once a collective hub has moved to a new location, may reduce the number of reissued abandon commands. Collective operators using the M_2 model abandoned the highest value target more frequently than IA operators. Further analysis is required to determine if the entire target icon must represent the target value, as was the case with the IA visualization, to be more salient. Opacity levels must also be validated to ensure an unique distinction between low-, medium-, and high-valued targets. Reiterating the task objective, to choose and move each collective to the highest value target for each decision, numerous times during training may also help mitigate operator mis*understanding*.

Target observations, which were additional target left-clicks that did not influence collective behavior or aid in accessing supplemental *information*, and interventions were additional undesired interactions. IA operators may have confused the target integer identifiers with the collective roman numeral identifiers causing additional target observations. Using distinct identifiers, such as integers versus letters, can potentially reduce the number of observations. IA operators' *capability* to identify objects far from their current attentional focal point may have been impeded by displaying all of the individual collective entities, collective and target icons, as well as the collective and target *information* pop-up windows. Asking *SA* probe questions about objects at various distances from the operator's current focal point is necessary to *understand* how clutter, or moving individual collective entities, may affect the operator's ability to identify the *SA* probe object of interest and answer the question correctly. The use of eye-tracking technology can provide improved insight regarding operator *understanding* and *usability* by recording where the operator was looking. *Understanding* what types of *information* the operator was potentially perceiving and comprehending, the difficulty of

identifying the desired *information* due to clutter, and the duration of *time* looking for *information* will illuminate why operators interacted with the system in a particular way.

The M_2 model enabled fewer commands, which was expected. Requiring operators to influence the decision-making process ensured better *performance*, because the operator was required to direct the decision-making process, versus operators who aided the decision-making process. Different strategies were used to fulfill the decision-making task; however, the most successful promoted more consensus decision-making (i.e., investigate commands) versus prohibiting exploration of particular targets (i.e., abandon commands). The *memorability* of both the models and visualizations enabled operators to refocus their attention on a previous action after answering the SA probe question, because of the required involvement of the operator (M_3 model) and established expectations of collective behaviors (M_2 model). The *predictability* of the M_3 model with the Collective visualization enabled operators to issue decide commands shortly after collectives were in a committed state. Collective operators using the M_3 model reported the best *control* mechanism responsiveness, which was anticipated due to the amount of operator influence and gained experience using the *control* mechanisms in the prior trial that used the M_2 model.

Transparency for human-collective systems can be achieved via different design strategies for specific system design elements and must be assessed holistically by *understanding* how the different factors impact transparency and are influenced by transparency. The four research questions assessed four categories of transparency factors that contribute to an *effective* system: (1) operator individual *capabilities*, (2) operator comprehension, (3) system *usability*, and (4) human-collective team *performance*. Ideal collective systems will enable operators with different individual *capabilities* to perform relatively the same, promote operator comprehension, be *usable*, and promote

high human-collective *performance*. As collective systems grow in complexity (e.g., size, heterogeneity), visualizations that show the individual collective entities will cause perceptual and comprehension challenges, as well as influence operator actions negatively. The same advantageous observation (i.e., dynamically seeing collective behaviors and support) from this analysis may not occur with large collectives (> 10000).

4.3 Visualization and Model with Visualization Conclusions

The Visualization Analysis in Chapter 4.1 evaluated transparency with respect to how the visualization impacted the human operators, operator comprehension, visualization usability, and human-collective performance. The Collective visualization was considered more transparent, because operators with different individual capabilities performed similarly in both the primary and secondary tasks, and the human-collective performance was higher compared to the IA visualization. The Model with Visualization Analysis in Chapter 4.2 considered how transparency embedded in both the models with the visualizations influenced the human operators, operator comprehension, system usability, and human-collective performance. The M_2 model with the Collective visualization combination did not support any of the research questions together, but did partially support specific research questions independently. Quantifying system transparency requires evaluating the transparency embedded in the various system design elements in order to determine how they interact with one another and influence human-collective interactions and performance. Designers of human-collective systems must build collective systems that are effective regardless of how heterogeneous or large the collective size may become, how simple or complex the collective behaviors are, and how challenging real-world use scenarios may be, such as bandwidth limita-

tions. Models (e.g., intelligent algorithms) that can aid operators to fulfill the sequential decision-making task that require operator influence and collective visualizations that are observable may be more resilient to real-world scenarios, and provide transparency to enable effective human-collective teams. The results of these two analyses will help inform design guidance for effective human-collective systems.

Chapter 5: Design Guidance for Human-Collective Systems

Design guideline recommendations were created in order to inform how transparency can be achieved for human-collective systems. The initial set of guidelines were inspired by the Visualization analysis [177] presented in Chapter 4.1, as well as the Model with Visualization analysis presented in Chapter 4.2 of the results collected from the single human-collective evaluations [29, 177]. The design guidance suggest recommendations with respect to visualizations, models (e.g., algorithms), and control mechanisms. Additional guidance are provided based on a review of the biological literature related to spatial swarms and colonies. The biologically-inspired design guidelines are categorized by seven biologically-inspired behaviors that cannot be investigated based on the results of the single human-collective evaluations, have not been investigated in depth by the existing literature, and can be explored in future human-collective evaluations.

The relationships between the design guidelines and the transparency factors in Figure 2.2 are discussed and emphasized using italics. Direct factors had immediate connections related to transparency, such as transparency impacts *performance*, whereas indirect factors typically influenced other factors, for example, *expectations* impact *control*. The relationships that are verbs between factors are also emphasized using italics. Some relationships have a positive influence, such as *promotes*, *fosters*, and *enhances*, on a particular factor, for example *effectiveness promotes usability*. *Impacts* may positively or negatively influence a factor, such as *workload impacts performance*. The final relationship described how *explainability communicates information*.

A discussion related to the limitations associated with limited or no communication, domain specific challenges (e.g., aerial or underwater), environmental challenges, and the type of collective systems used, is provided. The reliability of the guidelines inspired from the results and analysis of the single operator-collective evaluations and biological literature may be challenged by the identified limitations and must be further validated by future evaluations. Understanding how the limitations may impact the guidelines is needed in order to improve the robustness of the design guidance.

5.1 Design Guidance based on the Single Human-Collective Evaluations

The design guidelines inspired by the single human-collective evaluations' Visualization analysis (Chapter 4.1) as well as the Model with Visualization analysis (Chapter 4.2) are summarized in Tables 5.1 - 5.3. The recommendations are applicable irrespective of a visualization or model type. The presentation organization of these guidelines is in association with the visualizations, models, or control mechanisms.

5.1.1 Human-Collective Visualization Design Guidance

The research questions (R_1 - R_4) in Chapter 4.1 and their respective results and discussions are associated with design guidance related to visualization, which is summarized in Table 5.1. Design guidance that share a common idea, such as providing *information*, are discussed collectively, with the specific differences identified using the design guideline number. Several of the recommendations suggested providing particular types of *information*, such as collective behaviors (DG_1) and state (DG_2) *infor-*

mation, operator actions (e.g., issued commands) (DG_1), and system messages (DG_1), in order to facilitate operator *understanding*. The primary *information*, presented constantly throughout system usage, must be easily *observable* and comprehensible in order to maintain *SA*. The use of color can be used to distinguish objects from one another (DG_3), or to convey particular types of *information*, such as non-numeric values (DG_4) and command status (DG_5). Color coding can be useful for aiding *observability*, as long as the operator's cognitive capacity (DG_6) and *workload* are not exceeded. Other design

Table 5.1: Human-collective visualization design guidance.

DG_1. Provide *information* about the system and operator actions, such as the use of system messages and collective assignments windows.
DG_2. Provide *observable* collective state *information*, such as the use of color to denote different states.
DG_3. The use of colored borders is an *effective* method of distinguishing objects in the environment.
DG_4. The use of color and different opacity is an *effective* method of conveying a non-numeric value.
DG_5. Indicate the *status* of operator commands, such as a red indicator to denote completion of a command and green to denote an ongoing state.
DG_6. Limit the number of colors used to seven plus or minus two, which is consistent with human cognitive capacity (i.e., *capability*) [180].
DG_7. Provide detailed supplemental *information* to the operator, such as the use of *information* pop-up windows.
DG_8. Use distinct and unique identifiers for objects in the environment, such as integers versus letters.
DG_9. Provide a legend detailing *information* in order to alleviate memory demands of the operator.
DG_{10}. Provide *information* about collective behavior that coincides with operator mental models of operation, such as abandoning a target will result in zero individual collective entity support.
DG_{11}. Provide indicators that identify which particular objects are currently selected, such as the Collective and Target fields in the Collective Request area.
DG_{12}. Provide the *predicted* state of a collective, such as a dynamic moving border.

strategies, such as the use of patterns, may need to be considered if operators are color blind. Supplemental *information* that can be accessed when an operator deems necessary, such as *information* pop-up windows, can provide more detailed *information* (DG_7); however, the operator must not *rely* on this *information* in order to fulfill tasks.

Quick and easy *observability* of objects and items of interest in the visualization are necessary for task completion. Object identifiers, such as letters to represent collectives and integers for targets, will ensure distinction (DG_8) and help mitigate mis*understanding*. Operator *workload* can be reduced by providing aids, such as legends (DG_9), that detail particular identifier *information*, such as the meaning of different colored individual collective entities. Operator comprehension can be facilitated by ensuring that *information* about collective behavior coincides with an operator's mental models of operation (DG_{10}), such as when a collective is selected, the corresponding field shows the collective identifier (DG_{11}). Mismatched *expectations* between what the operator thinks the collective will do and what the collective actually does can lead to undesirable operator *usability* behaviors intended to compensate for the mismatch, which may influence negatively task *performance*. Projecting future collective state *information* (DG_{12}) can potentially mitigate mismatched *expectations* by aiding operators in *understanding* how current collective actions are leading to future collective behavior. Further investigations are needed in order to determine the *effectiveness* of the design recommendations, as well as the *information* available from underlying models that can impact what is visualized. *Understanding* how real-world scenarios, which may introduce bandwidth limitations or other challenging situations, that impact *information* latency and contribute to inaccurate collective state *information* is essential to ensure positive human-collective behaviors and to design a resilient transparent visualization.

5.1.2 Human-Collective Model Design Guidance

The research questions (R_5 - R_8) in Chapter 4.2 and their respective results were associated with design guidance related to models and visualizations. The design guidelines in relation to the models are summarized in Table 5.2. The underlying models influence what *information* is presented on the visualizations and how the operator interacts with the collectives. Determining what characteristics are ideal in the models is necessary to promote *effective* and *efficient* human-collective teams.

Table 5.2: Human-collective model design guidance.

DG_{13}. Use underlying intelligent models (e.g., sequential best-of-n decision-making) *capable* of fulfilling the task without operator influence.
DG_{14}. Ensure that the underlying intelligent models compensate for environmental biases and other influential factors on the collective *process*es.

The recommendations suggest to use intelligent models that are *capable* of fulfilling tasks without the need for operator influence (DG_{13}). Potential real-world collective use scenarios will require operators to conduct multiple tasks simultaneously, which may increase workload and distract (i.e., reduction of *SA*) operators from fulfilling tasks. Models that can aid operators by contributing towards task completion may help reduce *workload* and allow the operator *time* to attend to various tasks. Although the model may have the potential to fulfill tasks independently, operator influence is still beneficial, since the individual collective entities have limited knowledge about the overall collective's state and behaviors that are *observable* to the operator. Ensuring that the models compensate for influential factors, such as environmental bias (DG_{14}), are necessary to coincide with operator *expectations* of the system operation. Operators may have limited knowledge about how system and environmental factors contribute to particular collective responses. Training operators about these factors prior to system

usage and providing explanations during the task can calibrate operator *expectations*; however, the models can alleviate *workload* and reduce potential mis*understanding* by proactively compensating for these factors throughout system usage.

5.1.3 Human-Collective Control Mechanisms Design Guidance

The design guidelines in relation to the *control* mechanisms associated with research questions (R_5 - R_8) in Chapter 4.2 and their respective results are summarized in Table 5.3. The *control* mechanisms enable the operator to influence the collective decision-making *process*. Ideal interactions will positively influence collectives and improve human-collective *performance*. Determining what *control* mechanism characteristics are necessary will enable *effective* human-collective interactions.

Table 5.3: Human-collective control mechanisms design guidance.

DG_{15}. Provide *control* mechanisms that influence the collective consensus decision-making *process* positively, such as the investigate command.
DG_{16}. Provide *control* mechanisms that can undo negative influence, such as cancel assignment.
DG_{17}. Limit the use of decision-making *control* mechanism only after a particular certainty value, such as 30% support for a specific target.
DG_{18}. Limit the amount of times operators can issue particular commands, such as one time for the abandon or decide command.

The recommendations suggest to use *control* mechanisms that influence the collective consensus decision-making *process* positively, such as investigate commands (DG_{15}). The investigate commands helped build support for particular targets with little influence. Ten uncommitted entities (5% of the collective population) transitioned to the favoring state after receiving and acknowledging the investigate command. Operators who wanted to build support rapidly needed to commit multiple investigate

commands in a short period of *time*. Only influencing a small portion of the collective population enabled better decision-making. The collective was able to investigate other potentially higher value targets in case the operator was currently favoring a lower value target. The collective's *capability* to investigate and build support for other targets simultaneously ensured better human-collective task *performance*. *Control* mechanisms that are more definitive (e.g., persist immediately after issuing the command) and have negative influence, such as the abandon commands, must be designed with great caution in order to ensure *effectiveness* and avoid undesirable operator behaviors. There were instances when the IA and the Collective operators issued abandon commands repeatedly for the same target, although it is only required to be issued once, and when the highest value target was mistakenly abandoned. Further investigations are required to determine how to improve the *efficacy* of *control* mechanisms, such as abandon, that can negatively influence task completion. Providing *control* mechanisms that can undo negative influence (DG_{16}) are necessary in order to avoid persistent undesired behavior and to ensure high task *performance*. Limiting the number of times operators can issue particular commands (DG_{18}), such as abandoning a target once, or deciding to commit to a target once, can help mitigate misuse issues. Implementing other limitations, such as the ability to issue a decide command only after 30% of the collective supports a target (DG_{17}), can also improve *usability* and mitigate undesired operator behaviors.

Transparency for human-collective systems can be achieved via different design strategies for specific system design elements, such as the visualizations, models, and control mechanisms. Understanding how embedding transparency into various system design elements can be combined in order to promote transparency holistically is necessary to guarantee desired human-collective behaviors, and promote optimal human-collective team performance. A collective system using similar models (e.g., best-of-n)

and hub-based colonies designed using the provided guidelines can help promote better transparency and enable *effective* human-collective teams.

5.2 Biologically Inspired Design Guidelines

While the design guidance in Chapter 5.1 was based on the single-human collective evaluations, there is a significant opportunity to further develop design guidance related to transparency for human-collective teaming. The biologically-inspired design guidelines derived from a literature review of biological spatial swarms (Chapter 2.0.1.1) and colonies (Chapter 2.0.2.1) behaviors are provided in Tables 5.4 - 5.10. Each guideline is designed to promote transparency need and is related to a biological behavior or characteristic identified from the literature that is applicable to robotic systems and human-collective teams. The particular biological behaviors that inspired each generalizable behavior are provided in Chapters 5.2.1 - 5.2.7. The guidelines that contribute to each particular generalizable behavior are discussed within their respective chapter. The guideline discussions address how the guideline was formulated, how it relates to the generalizable behavior, into which system design elements it can be embedded, and what transparency factors will be impacted, or influence that recommendation.

5.2.1 Undesirable Emergent Behaviors

The first behavior was inspired by honeybee colonies that use consensus decision-making, which can result in undesirable emergent behaviors (e.g., behaviors that can impede task completion), such as split decisions for a best-of-n decision problem (e.g., new nest selection). When biological processes, such as best-of-n, are codified into mod-

els, mitigations of such undesirable behaviors may be necessary. Some undesirable behaviors may require terminating the current task in order to initiate a new process version of the model. Honeybee colonies that make a split decision resettle and debate further in order to arrive at a consensus [5]. Completing a decision, for both the biological and robotic systems, is necessary in order to mitigate safety issues. Undesirable behaviors are wasteful and can be dangerous. Honeybee colonies, for example, must make a single decision quickly, because they are vulnerable and exposed in the environment (i.e., outside of the nest) throughout the duration of the decision-making process. Systems that mimic biological behaviors must consider how to mitigate undesirable behaviors and the negative influence on human-collective task completion.

Table 5.4: Design guidance for undesirable behaviors.

DG_{19}. Provide a likelihood *prediction* of a known undesirable emergent behavior, such as a split decision, along with a *prediction* error to the operator.
DG_{20}. Provide engagement *prompts* to indicate to the operator when an unknown undesirable behavior appears to be emerging.
DG_{21}. Provide suggestions to the operator, such as issuing particular commands, in order to prevent or minimize undesirable behavior and maintain safety.
DG_{22}. Provide *feedback* to the operator about the environmental and system characteristics contributing to an undesirable emergent behavior.
DG_{23}. Embed a procedure into the model to mitigate the development of the undesirable behavior.

The design guidance is to inform operators when an undesirable behavior is emerging, identify what may be contributing to the undesirable behavior development, and provide suggestions regarding how the operator can mitigate the behavior from developing further. Strategies, such as training operators on how to use the system, are recommended in order to mitigate undesirable emergent behavior prior to usage.

5.2.1.1 Known Undesirable Behavior *Prediction* and Error

Providing an operator a likelihood *prediction* of a known (i.e., previously seen and recognizable) undesirable emergent behavior along with a *prediction* error is the first design guideline (DG_{19}) in Table 5.4. This guideline was formulated considering how in*efficient* (e.g., waste of time and resources) and un*satisfying* (e.g., detrimental to safety) it was when the honeybee colony's best-of-*n* decision-making *process* resulted in a split decision during the *process* of moving to a new nest site [5]. Providing *relevant predictive information* that can inform the operator about the emergence of a known undesirable behavior emerging can help improve *efficiency*, *timing*, mitigate the loss of resources devoted to an undesirable behavior, and improve the overall human-collective's task *performance*. However, the operator needs *information* related to how accurate the *prediction* is and what is the system's confidence in the provided *prediction*, such as a confidence interval. This *prediction* error will inform the operator about the system's confidence (i.e., *reliability*) in the *prediction*, as well as calibrate their *expectations* and *reliance* on the system appropriately, which promotes better system *usability*.

The design guideline can be embedded into the model, visualization, or both. The model can aggregate the *information* from the collective and compute a probability (e.g., likelihood) that a known undesired emergent behavior is developing. A threshold value derived from the *relevant* biological literature, or defined and refined based on robotic system data collections and evaluations, can be used to trigger sending a message to the visualization software that determines how to present the *information* to the operator.

The *information* presented via the visualization must be *observable*, *explainable*, and *understandable* by the operator. Providing a clear, succinct, and *legible* explanation indicating what type of undesirable behavior appears to be emerging in a particular collec-

tive, and the error likelihood percentage associated with the *prediction* can be embedded into the visualization. The presentation of the *information* can be conveyed using different colors to indicate that the likelihood percentage is large and must be addressed in a *timely* fashion. Patterns can be used in place of or in conjunction with the colors in order to accommodate operators who are color blind. Text messages can indicate what particular known undesirable behavior is emerging, such as "split"; however, the text must be easily *observable* (e.g., large easily visible letters) and *understandable* in order to be an *effective* implementation. Representative symbols or icons may be an *effective* alternative that minimize the use of text and can be easier to perceive quickly to represent specific known undesirable behaviors. *Understanding* what is the known undesirable behavior, which collective it is affecting, and how *reliable* the *prediction* is, will properly calibrate the operator's *expectations* of the collective's future behaviors, and can improve *SA*, *reliance* of the *information*, and provide insight about system *capabilities*. The *frequency* at which the messages are presented must be considered in order to mitigate *workload* issues that may arise if the *information* becomes a nuisance.

Providing the likelihood *prediction* of a known undesirable emergent behavior along with a *prediction* error will ideally *prompt* operators to proactively interact with the system in order to mitigate further development of the behavior. *Control* mechanisms must be provided in order to enable the operator's attempt to mitigate successfully a known undesirable behavior, such as a command that promotes cohesion in order to avoid a physical split that may compromise the safety of the group. Training prior to system usage can help operators develop strategies to mitigate particular known undesirable behaviors. Additional suggestions about what commands can be issued can be provided by the system (DG_{21}), are discussed in Chapter 5.2.1.3.

Providing the *prediction* and associated error of a known undesirable behavior emerging promotes transparency by improving the system's *usability*. The transparency promoted in this design guideline needs to provide useful and *explainable information* that allows the operator to *understand* the collective's future *state*. This level of transparency can promote better *SA*, which prompts the operator to take appropriate preventative actions that result in improved human-collective *performance*.

5.2.1.2 Engagement *Prompt* for Unknown Undesirable Emergent Behavior

Unlike known undesirable emergent behaviors that can be anticipated prior to system usage and have recognizable characteristics, unknown emergent behaviors are known to occur with collective systems and cannot be anticipated or recognized. The use of collective systems in challenging environments has been proposed due to their *adaptability*; however, anticipating behaviors that may arise in those environments is challenging and often unattainable. Providing engagement *prompts*, such as indicators or warnings, to the operators when an unknown undesirable behavior is potentially emerging, DG_{20} in Table 5.4, is imperative in order to maintain *SA* and attract the operator's attention to *information* indicating that a potential unknown emergent behavior is occurring that may impede task completion. This guideline was formulated considering what alternatives were needed when an unknown undesirable emergent behavior occurred, unlike situations in which a known undesirable emergent behavior exists, DG_{19}.

The design guideline can be embedded into the model, visualization, or both. The model can estimate the collectives' *state* and its deviation from the expected or desired

state. The percent deviation from task completion is an example of tracking the influence of undesirable behaviors in general. A threshold value derived using DG_{19} principles from Chapter 5.2.1.1, can be used to trigger an analysis process in the visualization software that determines how to present the *information* to the operator.

Information presented on the visualization can *promote understanding*, as long as it is *observable*. Providing a clear and *legible* engagement *prompt* of the collective's deviation between the current *state* and the expected or planned *state* can be embedded into the visualization. Showing the percent deviation from task completion in a particular collective can be conveyed as a numerical percentage value outlined using bold lines in order to attract the operator's attention to the relevant *information* pertaining to an unknown undesirable emergent behavior that is negatively influencing a particular collective's task *performance*. The engagement *prompt* can remain visible until the percent deviation drops below the threshold. Attracting the operator's attention by using multimodal strategies (i.e., visual, auditory, and tactile), to the percent deviation value can calibrate the operator's *expectations* of the collective's future behaviors, and provide insight that helps build mental models of the system's *capabilities*. Informing operators that the system has fewer *capabilities* may negatively impact its *credibility* and perceived *reliability*, if the operator's mental model indicated that the system was *capable* of mitigating or informing the operator about the undesirable behavior. Training prior to system usage can help properly calibrate the operator's *expectations* about system *capabilities* and how much they can *rely* on the system. Operational environment characteristics will influence which multimodal strategies will best attract operators' attention to the *relevant information*. Attracting attention to the collective's deviation from the current *state* to the expected or planned *state* has positive and negative implications for *SA*. The operator's *SA* can improve by knowing that something is impeding the

human-collective team's *capability* to fulfill a task; however, if the engagement *prompt* is highly salient the operator may become distracted and not attend to high priority tasks associated with the system's collective(s), which may contribute to further deviation from task completion due to undesirable behavior development. *Understanding* how to maintain attention across multiple collectives simultaneously is necessary in order to promote a usable human-collective system.

Providing an indicator that an unknown undesirable emergent behavior is impeding task completion will ideally prompt operators to proactively interact with the system in order to halt or mitigate further development of the undesirable behavior, or contribute to achieving a prior desirable *state*. Different types of *control* mechanisms must be provided in order to help mitigate the undesirable behavior. Recording which *control* mechanisms have been used is necessary in order for the operator to *understand* whether using specific *control* mechanisms mitigated the behavior or contributed to the undesirable behavior's further development. Training prior to system usage can help operators develop strategies to recognize the influence of *control* mechanisms on collectives and develop mental models that can help calibrate neglect benevolence, which is the *time* required to allow a system to stabilize before issuing new commands [136].

Providing engagement *prompts* when an unknown undesirable behavior is potentially emerging promotes transparency by improving the system's *usability*. The transparency promoted in this design guideline, similar to DG_{19} from Chapter 5.2.1.1, needs to provide *explainable information* that allows the operator to *understand* the collective's future *state* with respect to the task goal. This level of transparency can promote better *SA*, which prompts the operator to inquire about what characteristics are contributing to collective behavior changes, such as using feedback from Chapter 5.2.1.1, and taking appropriate preventative actions that result in improved human-collective *performance*.

5.2.1.3 Suggestions to Mitigate Undesirable Behavior

Providing suggestions to the operator about what actions (e.g., control mechanisms) can be used to mitigate the further development of a known or unknown undesirable emergent behavior is DG_{21} in Table 5.4. Providing suggestions can alleviate the *workload* associated with determining how to prevent or minimize further development of the undesirable emergent behavior. This guideline was formulated considering how the system can proactively aid the operator in the effort to prevent or mitigate further development of the undesirable emergent behavior. Providing *relevant* suggestions can help improve *efficiency*, *timing*, mitigate loss of resources devoted to an undesirable behavior, and improve the overall human-collective's task *performance*. However, human-collective system designers must consider how abiding by the system suggestions effects operator *expectations*. Some operators may expect an immediate decrease in the further development of the undesirable behavior, although time is needed for the collective behavior to stabilize (i.e., neglect benevolence), and if that *expectation* is not *satisfied*, it may cause undesirable operator behaviors. The *credibility*, perceived *reliability*, and *reliance* of the system may decrease as a consequence of the misaligned *expectations*, which may cause the operator to take *control* of all proceeding collective behaviors, negatively impacting system *usability*. Providing the collectives' *state* and deviation from the expected or desired *state information*, similar to the example from Chapter 5.2.1.2, as well as develop accurate mental models about system response *times* during training prior to system usage may mitigate operator *expectation* misalignment issues.

The design guideline can be embedded into the model, visualization, or both. The model can initially estimate the collectives' current *state* and its deviation from the expected or desired *state* by implementing DG_{20} from Chapter 5.2.1.2. A threshold value

derived using DG_{19} principles from Chapter 5.2.1.1, can be used to trigger a decision support tool or predictive simulation tool, which can calculate what actions, such as *control* mechanism, will contribute to a lower deviation from the expect or desired *state*. Once the model has determined the best suggestion can be communicated to the visualization software that determines how to present the *information* to the operator.

The *information* presented on the visualization must be *observable* and *explainable* in order to *foster understanding* of the longer term implications of the potential mitigation alternatives. Providing a clear, succinct, and *legible* suggestion indicating what action the system recommends and why for a particular collective can be embedded into the visualization. The presentation of the *information* can be conveyed on a pop-up window near the *control* mechanisms. Options can be provided to the operator to "accept" or "cancel" the system's recommendation. A text message can indicate what particular *control* mechanism, for example, "abandon" search for a target, is suggested; however, the text must be easily *observable* (e.g., large easily visible letters) and *understandable* in order to be an *effective* implementation. Alternatively, the suggested *control* mechanism interactive icon can be outlined using bold lines in order to attract the operator's attention to the particular *control* mechanism, potentially improving *SA*. The window can remain visible to the operator until either the suggestion is accepted, canceled, or the *time* to issue that particular suggestion expires.

Providing suggestions about what operator actions can mitigate the further development of a known or unknown undesirable emergent behavior will ideally *prompt* operators to interact with the system. *Control* mechanisms must be provided that enable the operator's attempt to mitigate an undesirable behavior, and the *capability* to adjust software parameters used to determine mitigations, as adjustments to the parameters can result in different mitigation suggestions on the collective's goal achievement out-

comes. Recording which *control* mechanisms have been used is necessary in order for the operator to *understand* whether using the *control* mechanism mitigated the behavior. Training prior to system usage can help develop operator mental models regarding the system *usability* and how to interact with the recommended suggestion.

Providing suggestions to the operator about what actions can mitigate the further development of a known or unknown undesirable emergent behavior promotes transparency by improving the system's *usability* and human-collective *performance*. The transparency promoted in this design guideline needs to provide *explainable information* that allows the operator to *understand* what actions can mitigate further development of undesirable emergent behavior and provide *control* mechanisms to execute such actions. This level of transparency can alleviate *workload* by promoting better *SA*.

5.2.1.4 *Feedback* about Environment or System Characteristics

Relevant feedback can be provided to the operator in order to provide *context* regarding what environmental and system characteristics are contributing to a known or unknown undesirable emergent behavior and *justify* why the behavior is occurring, guideline DG_{22} in Table 5.4. Providing *feedback* promotes *explainability* by being *learnable*, which can improve operator *satisfaction* and *SA*, as well as calibrate operator *expectations* regarding the system's *capability* limitations. This guideline was formulated considering how operators often do not *understand* why collectives are behaving in a particular manner. Providing *feedback* is a useful method for promoting transparency; however, too much *feedback* may distract operators and can cause higher *workload*.

The design guideline can be embedded into the model, visualization, or both. The model can add environment (e.g., obstructions in the environment) and system charac-

teristic (e.g., multiple solution options being weighted equally) identifiers to the *information* provided from the collective prior to aggregating it and computing a probability that a known undesired emergent behavior is developing, similar to Chapter 5.2.1.1 example. Errors (i.e., failure to classify *information*) in the model logic may trigger the initiation of other procedures when the model tries to compute the probability for an unknown undesired emergent behavior. Recording the associated errors and what *information* was missing in order to classify the behavior as a known undesirable emergent behavior must be provided to the operator. A threshold value derived using DG_{19} principles from Chapter 5.2.1.1, can be used to trigger sending a message to the visualization software that determines how to present the *feedback* to the operator.

The *feedback information* presented on the visualization must be *explainable* in order to *foster understanding* regarding what type of known or unknown undesirable behavior appears to be emerging in a particular collective, including what type of characteristics, misalignment of the current *state* to goal *state*, or missing *information* is contributing to the development of that behavior. The *feedback information* can be presented in various ways, including color coding, text messages, representative symbols, or icons that were discussed in Chapter 5.2.1.1. The identification of what environmental or system characteristic is contributing to the development of the known undesirable emergent behavior can be presented as a representative icon in order to mitigate the amount of text provided to the operator. *Understanding* highly detailed *feedback* associated with unknown undesired emergent behaviors will be more challenging and may require supplemental text in order to ensure the operator *understands* what *information* is missing. Designers must balance how *information* is presented via color coding, text, and representative symbols or icons. Using too many colors and symbols or icons may exceed operators' cognitive *capabilities*. *Understanding* what the known undesirable emergent behavior is,

what *information* is missing from the model logic in order to classify a behavior as a known undesired emergent behavior, which collective it is affecting, and what environmental or system characteristic is contributing to the behavior development will help properly calibrate the operator's *expectations* of the collective's future behaviors, and can improve *SA*, the *reliance* of the *information*, and provide insight about system *capabilities*. The *feedback* can be accessed as supplementary *information* in order to mitigate *workload* issues that may arise with too much supplied *information*.

Providing *feedback* to the operator will not necessarily *prompt* operators to proactively interact with the system; thus, *control* mechanisms that enable an operator to access the *feedback information* must be provided.

Providing *feedback* about what *information*, or missing *information*, appear to be contributing to the development of a known or unknown undesirable emergent behavior promotes transparency by improving the system's *explainability*, which fosters the operator's *understanding*. The transparency promoted in this design guideline needs to provide *explainable information* that allows the operator to *understand* why the development of a known or unknown emergent behavior is occurring and what characteristics, or missing *information*, are contributing to the development. This level of transparency can improve *SA* and provide accurate *justifications* for operator actions.

5.2.1.5 Undesirable Behavior Mitigation Procedure

Embedding procedures into the model *capable* of mitigating the development of known undesirable emergent behaviors is design guideline DG_{23} in Table 5.4. Providing the system *capability* to mitigate known undesirable emergent behaviors can be more *effective* and *efficient*, which *promote* better *usability*. This guideline was formulated consider-

ing how system *control* (e.g., autonomy) can help alleviate operator *workload* associated with completing tasks. Designers of human-collective systems must consider how to maximize the strengths of both the model and operator in order to promote optimal *performance*. Using a model designed to achieve a task without operator influence (e.g., a best-of-*n* model for highest value target identification task) will aid operators and can improve human-collective performance.

This design guideline is intended to be embedded into the model only. The model can use similar *information* aggregation, probability computation, threshold triggers, and decision support or *predictive* simulation tools discussed in Chapters 5.2.1.1 and 5.2.1.3 in order to formulate a *process* of mitigating known undesirable emergent behaviors. The decision support and *predictive* simulation tools as those can help determine the best set of actions the model can take and execute those actions accordingly. The known undesired emergent behavior mitigation *process* may experience an error or fail to proceed forward if the behavior is unknown. A separate *process* must be implemented in order to accommodate for unknown undesired emergent behaviors. The system may return to a previous *state* before the unknown undesired behavior began to emerge in order to implement another strategy to mitigate the undesired behavior, or may try other sub-optimal actions to determine if those actions contribute to attaining a proper mitigation. Both *processes* will be iterative, as the system will constantly use the collective *information* to monitor for known undesirable emergent behaviors and follow the respective *process* to mitigate further development of the behavior, depending on whether it is known or unknown.

Providing *information* on the visualization about the system's *capability* to mitigate known undesirable behaviors is unnecessary; however, training prior to system usage is needed in order to promote accurate mental models of system *reliability* and *under-*

standing regarding system strategies when an unknown undesired behavior emerges. Depending on the *reliability* or strategy, an operator may be prompted to influence the system by providing additional *information* to influence the system decision-making outcome or overrule the system's decision depending on the current situation.

Embedding procedures into the model *capable* of mitigating the development of known undesirable emergent behaviors and strategies to cope with unknown undesired emergent behaviors promotes transparency by improving the system's *usability*. This level of transparency can alleviate operator *workload*, while promoting better human-collective *performance*. The model will have *control* over determining how to mitigate or cope with undesired emergent behaviors; however, operators will be *capable* of supplementing *information* to improve decision-making or override the model.

5.2.2 Cohesion

Cohesion is the degree of connectedness in a group and is the second biologically inspired behavior. The most common benefit of cohesion in both biological and robotic collectives is increased safety of the individual collective entities by being a part of a collective group. Honeybees, fish, and birds maintain cohesion in order to reduce the number of isolated individuals or small groups of members from attacks made by predators [6, 14] or environmental factors [5]. Robotic individual collective entities will experience similar challenges associated with the environment, and may encounter adverse individuals depending on the situation. Cohesive groups can achieve complex geometries that are beneficial to maneuver around objects in an environment, evade adverse individuals, and provide the capability for individuals to access resources, as is observed in tuna that use parabolic formations for cooperative hunting [16]. Systems

that mimic cohesive biological behaviors must consider how to promote cohesion.

The design guidance is to inform operators of the current collectives' cohesion status, predictive cohesion information about the collectives, and feedback about why the collectives' cohesion is or is not changing. Operators who have a better understanding of current and predicted cohesion states, as well as an understanding of why the state is or is not changing will be able to promote better cohesion due to improved SA.

Table 5.5: Design guidance for cohesion.

DG_{24}. Provide current *status* and *predictive information* to the operator about the collective's cohesion with respect to the given task, system *state*, and environment *state*, such as percentage of aggregation.
DG_{25}. Provide *feedback* to the operator regarding why the collectives' cohesion is or is not changing.

5.2.2.1 Current *Status* and *Prediction* of Cohesion

Providing the operator a collectives' current *status* and *predictive information* about the collective's future cohesion with respect to the given task, as well as the system and environment *state* is design guideline DG_{24} in Table 5.5. This guideline was formulated considering how challenging it may be for operators to *observe* and *understand* whether the cohesion of a collective is within a desired range. Varying geometry will cause density changes that will challenge the operator's *capability* to determine whether the behavior is positive or negative. Providing the current *status* of the collective will attract the operator's attention to *relevant information*, improving their *SA*, and reducing the *workload* associated with determining what is happening with cohesion. *Relevant predictive information* about the collective cohesion will provide insight and promote *learnability* about future collective cohesion.

The design guideline can be embedded into the model, visualization, or both. The model can aggregate the available *information* from the collective and compute a current cohesion *status*, such as percent cohesion value, with respect to the given task, as well as the system and environment *state*. Based on the influence of the task, system, and environment *states* a *prediction* can be computed to determine the collective's future *state*. A threshold value can be used to trigger sending a message to the visualization software that determines how to present the *information* to the operator. The number of seconds a future *state* is projected and the trigger value can be derived using DG_{19} principles from Chapter 5.2.1.1.

The *information* presented to the operator must be *observable* and *explainable* in order to *foster understanding*. Providing a clear, succinct, and *legible* current and *predicted* cohesion *state* message can be embedded into the visualization. The presentation of the *information* can be conveyed as a numerical percentage value with a header to indicate which percentage value it is, for example, "current" versus "predicted". Using a numerical percentage value suggests that the operator has prior knowledge, potentially from training, about what cohesion percentage values are ideal. Designers can use different colors, patterns, or representative symbols to indicate whether the current and predicted cohesion *states* are good or bad. A check mark, for example, can be used to indicate good, and a "x" cross can indicate bad. Representative symbols or icons may be an *effective* alternative for operators who have a color vision deficiency, minimizes the use of text, which can impose more *workload* on an operator, and can be easier to *observe* quickly. The presentation of the current and future cohesion *state information* can be accessed in supplemental windows in order to reduce clutter on the visualization; however, if the *information* is necessary to successfully complete a task, the *information* must be presented on the visualization near the respective collectives.

Providing the current and predicted cohesion *state information* will ideally prompt operators to proactively interact with the system in order to ensure cohesion is in a desirable range. *Control* mechanisms must be provided in order to enable the operator's attempt to increase cohesion, such as a command that promotes cohesion in order to avoid a physical split that may compromise the safety of the group. Recording which *control* mechanisms have been used is necessary in order for the operator to *understand* whether using the *control* mechanism promoted better cohesion. Training prior to system usage can help operators develop accurate mental models of how particular *control* mechanisms influence collective cohesion.

Providing the current *status* and *predictive information* about the collective's future cohesion with respect to the given task, as well as the system and environment *state* promotes transparency by improving the system's *usability*. The transparency promoted in this design guideline needs to provide *observable* and *explainable information* to *understand* the collectives' current and future cohesion *state*. This level of transparency can promote better *SA*, which prompts the operator to take appropriate preventative actions that result in improved human-collective *performance*.

5.2.2.2 *Feedback* about Cohesion *State*

Providing *relevant feedback* to the operator regarding why the collectives' cohesion is or is not changing can supply *context* and *justification* for the collective's behavior, guideline DG_{25} in Table 5.5. Providing *feedback* promotes *explainability*, which fosters *understanding*, by being *learnable*. Operator *satisfaction* and *SA* can be improved, resulting in more accurate calibrations of operator *expectations* regarding the system's *capability* limitations. This guideline was formulated considering how operators often do not

understand why collectives are or are not behaving in a particular manner. Providing *feedback* is a useful method for promoting transparency; however, providing too much *feedback* may distract operators and can cause higher *workload*.

The design guideline can be embedded into the model, visualization, or both. The model can add task (e.g., changing task priority), system (e.g., multiple solution options being weighed equally), or environment (e.g., obstructions in the environment) characteristic identifiers to the *information* provided from the collective prior to aggregation and compute the current cohesion *status*, similar to DG_{24} in Table 5.5. The rate of cohesion change can be computed using the current cohesion *state* and the prior recorded *state information* in order to determine whether the collective's cohesion is or is not changing. The identifiers from the current and previous cohesion states can be compared, and those that are different can indicate what characteristics may be contributing to the cohesion behavior change. A threshold value derived using DG_{19} principles from Chapter 5.2.1.1, can be used to trigger sending a message to the visualization software that determines how to present the feedback to the operator.

The *feedback information* presented on the visualization must be *explainable* in order to *foster understanding* regarding why the collective cohesion is or is not changing and what characteristics may be influencing the behavior change. The *feedback information* can be presented in various ways, including color coding, text messages, representative symbols, or icons that were discussed in Chapter 5.2.7.2. Representative icons can be used to identify whether the collective cohesion is changing and in what direction, such as an upward facing arrow to indicate positive cohesion change. No change in collective cohesion status can be represented using a symbol, such as an "x" indicator. The identification of what characteristic is contributing or impeding cohesion *state* change can also be presented as a representative icon; however, other strategies may become

more useful if too many symbols are being used and the operator's cognitive *capability* is exceeded. *Understanding* how collective cohesion is or is not changing and what characteristics may contribute to that change will help calibrate the operator's *expectations* of the collective's future behaviors by providing *justifications* for change, which will improve *SA*. The *feedback* can be accessed as supplementary *information* in order to mitigate *workload* issues that may arise when too much *information* is supplied.

Providing feedback to the operator will not necessarily prompt operators to proactively interact with the system; thus, control mechanisms that enable an operator to access the feedback *information* must be provided.

Providing *relevant feedback* regarding why the collectives' cohesion is or is not changing and what characteristics may be contributing to that cohesion *state* change promotes transparency by improving the system's *explainability*, which fosters the operator's *understanding*. This level of transparency can improve *SA* and provide accurate *justifications* for collective cohesion *state* changes.

5.2.3 *Timing* to Maintain Cohesion

The timing of individual collective entity behaviors is critical to maintain cohesion of a collective and must be considered. This behavior was inspired after identifying that the rate at which members of cohesive biological groups exchange positions is crucial in order to maintain long-range cohesive order [17]. Reshuffling too quickly can have detrimental effects on the collective and increase safety risks. Robotic collective systems that mimic cohesive biological behaviors must consider timing. Details about the design guideline associated with this behavior are provided in Chapter 5.2.3.1.

Table 5.6: Design guidance for timing to maintain cohesion.

DG_{26}. Provide suggestions to the operator, such as issuing particular commands, at a specific *time* (e.g., to accommodate for neglect benevolence) in order to improve a collective's cohesion.

5.2.3.1 Suggestions to Improve Cohesion

Providing suggestions to the operator regarding what potential actions that can be taken at a specific *time* to improve a collectives' cohesion is DG_{26} in Table 5.6. Providing the operator with suggestions can alleviate the *workload* associated with determining what actions will improve cohesion. Providing the specific *time* when an action must be taken can improve the efficacy of the human-collective team and the *effectiveness* of the action. This guideline was formulated considering how the system can proactively aid the operator in promoting better collective cohesion. Providing *relevant* suggestions and when those suggestions must be implemented can help improve *efficiency*, *timing*, and improve the overall human-collective's task *performance*. Repeated interactions with the system and what suggestions are recommended at specific *times* during training and system usage can promote *learnability*. The operators can *learn* what actions are necessary for particular known emergent behaviors, strategies that can be useful during unknown emergent behaviors, and anticipate when those respective actions must be taken. Human-collective system designers will need to consider how operator *expectations* may be influenced by following the system recommendations. Some operators may expect an immediate improvement of collective cohesion, although *time* is needed for the collective behavior to stabilize (i.e., neglect benevolence). *Expectations* that are not *satisfied* may reduce system *credibility*, perceived *reliability*, and *reliance*. The operator may take *control* of all proceeding collective behaviors, negatively impacting system

usability. Accurate mental models about system response *times* during training prior to system usage may mitigate operator *expectation* misalignment issues.

The design guideline can be embedded into the model, visualization, or both. The model can use the principles from DG_{24} in Table 5.5 to compute a current cohesion *status*. A decision support tool or predictive simulation tool, similar to that from DG_{21} in Table 5.4, can calculate what actions, such as *control* mechanisms, will contribute to a higher future collective cohesive *state*. The time to execute the suggested action must be sufficient in order to convey the message to the operator and allow time for the operator to comprehend the message. A threshold value derived using DG_{19} principles from Chapter 5.2.1.1 can be used to trigger sending the message to the visualization software that determines how to present the *information* to the operator.

The *information* presented on the visualization must be *observable* and *explainable* in order to *foster understanding*. Providing a clear, succinct, and *legible* suggestion indicating what actions are recommended to improve cohesion for a particular collective and when those actions must be taken can be embedded into the visualization. The presentation of the *information* can be conveyed on a pop-up window near the *control* mechanisms. Options, similar to those mentioned in DG_{21} in Table 5.4, can be provided to the operator to "accept" or "cancel" the system's recommendation. A text message can indicate what particular *control* mechanism, for example, "abandon" search for a target, is suggested. A timer countdown representative icon must be provided near the recommended action in order to indicate when the operator must make the suggested action. The pop-up window can become visible to the operator when a suggested action is recommended and remain visible to the operator until either the suggestion is accepted, canceled, or the *time* to issue that particular suggestion expires.

Providing suggestions about what operator actions can improve a collective's cohesion and when a particular action must be taken will ideally prompt operators to interact with the system. *Control* mechanisms must be provided that enable the operator's attempt to improve cohesion, and the capability to adjust software parameters used to determine which actions are ideal, as adjustments to the parameters can result in different action suggestions. Recording which *control* mechanisms have been used is necessary in order for the operator to *understand* whether using the *control* mechanism promoted better collective cohesion.

Providing suggestions to the operator about what actions can improve collective cohesion and the timing to issue the recommended actions promotes transparency by improving the system's *usability*, via *efficiency* and *effectiveness*, which ultimately improve human-collective *performance*. The transparency promoted in this design guideline needs to provide *explainable information* that allows the operator to *understand* what actions can improve collective cohesion and when those actions must be issued. *Control* mechanisms to execute such actions must be provided to the operator. This level of transparency can alleviate *workload* and promote better *SA*.

5.2.4 Individual Collective Entities Roles

Biological individual collective entities have roles that can persist or change depending on various characteristics. Physiological characteristics, such as body length [18], nutritional state [19], and age [5, 48] can influence particular roles and behaviors. Fish body length [18] and nutritional state [19] can determine placement within a group, while honeybee and ant [5, 48] age can impact particular roles. A general pattern of honeybee and ant role change starts where younger workers remain inside the nest to serve as

nurse or honeycomb builder bees and as they age, transition to foraging and scout bee roles outside the nest [5]. Robotic collective systems can assign particular individual entities specific roles based on similar "physiological" characteristics, such as software or hardware features. Environmental characteristics can also influence roles and behaviors, such as the inside of a honeybee colony's hive (congestion of the adult bees, numerous immature bees, and expanding food reserves) and outside the hive (plentiful forage in the spring time) have been correlated with starting the process of queen rearing [5]. Robotic individual collective entities who possess information inside and outside of a hub, similar to those exemplified by the scout honeybees, can initiate processes. Mimicking biological roles and characteristics to select individuals for specific roles can be used in robotic human-collective systems.

Table 5.7: Design guidance for individual collective entities roles.

DG_{27}. Provide engagement *prompts* to the operator if the number of individuals in particular roles decreases below a critical threshold.
DG_{28}. Provide suggestions about how to transition individual entities into new roles in order to avoid falling below a critical threshold.
DG_{29}. Provide *feedback* to the operator about why *capabilities* or roles are changing.
DG_{30}. Implement model strategies to re-assign individuals to new roles if other members fall below critical *capabilities*, such as low battery power.

The design guidance for the roles behavior is to provide prompts to operators if characteristics fall below critical thresholds, suggestions or strategies on how to transition individuals into new roles, and feedback regarding why individual collective entity capabilities or roles are changing. Using heterogeneous collectives can maximize the strengths of the operator and the different individual collective entities in order to promote high human-collective performance.

5.2.4.1 Role Engagement *Prompts*

Collective roles can be distributed among individual collective entities in order to ensure task completion. Circumstances may arise when the number of individual collective entities performing a particular role falls below a critical threshold and inhibits task progression. Various types of behaviors, including known or unknown emergent behaviors, as well as hardware or software changes in the individual collective entities may cause issues with task progression. Providing engagement *prompts*, such as indicators or warnings, to the operators when the critical threshold has been reached is DG_{27} in Table 5.7. Engagement *prompts* can be used to draw attention to changes that will impact role fulfillment and can calibrate operator *expectations* accurately, maintain *SA*, and *prompt* the operators to take actions to maintain task progression. This guideline was formulated considering what alternatives are needed when circumstances arise that cause individual collective entities to no longer fulfill a particular role that is needed to fulfill the mission objectives.

The design guideline can be embedded into the model, visualization, or both. A set of roles, and the optimal number of individual collective entities with particular *capabilities* to perform a respective role can be defined and associated with tasks. Identifiers can distinguish the *capabilities* of the individual collective entities, as well as what role the entity is performing. The *information* from the collective can be aggregated together in order to compute a role *status* relative to task completion. The *status information* can be used to determine whether the collective is progressing toward task completion. Missing *information* due to loss of individual collective entities, which may occur due to environmental impacts, loss in individual entity *capabilities*, a critical sensor or actuator, or imperfect communication, are examples of characteristics that may influence no task

progression. A threshold value derived using DG_{19} principles from Chapter 5.2.1.1 can be used to trigger an analysis process in the visualization software that determines how to present the *information* to the operator.

Engagement *prompts* presented on the visualization can *promote understanding*, as long as they are *observable* and *explainable*. Providing a salient, clear, and *legible* engagement *prompt* can be embedded into the visualization indicating that the number of individual collective entities performing a particular role has dropped below an acceptable threshold. The number of individual collective entities performing a particular role out of the total number of entities needed to complete a task can be presented on an *information* pop-up window as a numerical ratio message, such as "Foraging: 18 out of 20". The engagement *prompt* can remain visible until the number of individual collective entities is no longer below the critical threshold. Attracting the operator's attention by using non-visual multimodal strategies (e.g., auditory or tactile) can expedite detection of the engagement *prompt* and alleviate *workload*, as long as the operational environment enables proper detection of the cues. The operator's *SA* can improve by knowing the role *status*, as long as the engagement *prompt* is not too salient. Highly salient engagement prompts may distract operators away from attending to other high priority tasks associated with the system's collective(s), which may contribute to further reduction of individual collective entities needed for particular roles. *Understanding* how to maintain attention across multiple collectives simultaneously is necessary in order to promote a usable human-collective system.

Providing a *prompt* that indicates the number of individual collective entities in a particular role has fallen below a critical threshold will ideally cause operators to interact proactively with the system in order to increase the number of entities in that role. Different types of *control* mechanisms must be provided in order to help increase

the number of individual collective entities performing a role or to mitigate the lose of entities due to specific circumstances. Recording which *control* mechanisms have been used is necessary in order for the operator to *understand* whether using particular *control* mechanisms maintained, decreased, or increased the number of individual collective entities fulfilling a role. Training prior to system usage can help operators develop strategies to recognize the influence of *control* mechanisms on individual collective entities and develop mental models that can help calibrate neglect benevolence, since the collective role behavior may need time to stabilize.

Engagement *prompts* indicating the *status* of the number of individual collective entities performing specific roles below an acceptable and critical threshold promotes transparency by improving the system's *usability*. The transparency promoted in this design guideline, similar to DG_{19} from Chapter 5.2.1.1, needs to provide *explainable information* that allows the operator to *understand* the collective's role *status* with respect to the task progression. This level of transparency can promote better *SA* and calibrate operator actions in order to improve human-collective *performance*.

5.2.4.2 Suggestions to Transition Individual Collective Entities into New Roles

Providing suggestions to the operator about how to transition individual collective entities into new roles in order to avoid falling below a critical threshold of entities performing a role is DG_{28} in Table 5.7. This guideline was formulated considering how the system can proactively aid the operator with determining how to transition individual collective entities into new roles, which can alleviate *workload*, as well as improve

SA and system *usability*. Providing *relevant* suggestions can help improve *efficiency*, *effectiveness*, *timing*, mitigate loss of resources devoted to determining how to transition individual collective entities into new roles, and improve the overall human-collective's task *performance*. Operator *expectations* may be negatively impacted by providing suggestions if the operator becomes overly reliant on the system information in order to fulfill a task or it distracts the operator for other higher priority tasks. Some operators may expect an immediate response of individual collective entities transitioning into new roles, although time is needed for the collective behavior to stabilize (e.g., neglect benevolence). Misalignment in operator *expectations* of system *usability* may also cause dis*satisfaction*, reducing system *credibility* and perceived *reliability*. The operator may take *control* of all proceeding individual collective entity role transitions, negatively impacting system *usability* and potentially reducing human-collective *performance*. Accurate mental models about system response *times* during training prior to system usage may mitigate operator *expectation* misalignment issues.

The design guideline can be embedded into the model, visualization, or both. The principles about establishing a set of roles, number of individual collective entities needed for particular roles, distinguishing individual collective entities' *capabilities*, and what roles the entities are currently performing from DG_{27} in Table 5.7 can be encoded into the model. A probability, similar to DG_{19} from Chapter 5.2.1.1, can be computed to determine whether current actions are influencing task progress negatively and if the critical threshold will be reached relatively soon. A minimum time remaining estimate to take preventative actions on role transitions, using the DG_{19} principles from Chapter 5.2.1.1, can trigger a process of determining what actions are needed to transition individual collective entities into new roles, which have the necessary *capabilities*. A decision support tool or *predictive* simulation tool, as cited for DG_{21} in Chapter 5.2.1.3 can

identify the actions needed to make *efficient* and *effective* individual collective entity role transitions. Once the model has determined the best suggestion it can be communicated to the visualization software that determines how to present the *information*.

The *information* presented must be *observable* and *explainable* in order to *foster understanding* of the longer term implications of the potential suggestions. Providing a clear, succinct, and *legible* suggestion indicating which individual collective entities can be transitioned into a new role, how to make that transition, and when the transition must be taken can be embedded into the visualization. A message presented on a pop-up window near a respective collective can identify the suggestion. The subgroup of individual collective entities that can transition to a new role can be presented using text, such as "Worker". The role that the individual collective entities can transition into can be represented by an arrow pointing towards the new role written in text, such as " $\Longrightarrow$ Forager". Below the current role transition into a new role *information*, a second line of *information* can indicate the recommended action to transition roles and a numerical *time* counting down when the action can be taken. Using a combination of text and representative symbols will alleviate the *workload* associated with reading the system provided message. The window can remain visible to the operator until either the suggestion is accepted, canceled, or the *time* to issue that particular suggestion expires.

Providing suggestions about what actions can help transition individual collective entities into new roles will ideally *prompt* operators to interact with the system. *Control* mechanisms must be provided that enable the operator's attempt to transition individual collective entities into new roles, and the *capability* to adjust software parameters used to determine the suggestions, as adjustments to the parameters can result in different suggestions based on the collective's goal achievement outcomes. Recording which *control* mechanisms have been used is necessary in order for the operator to *understand*

whether using the *control* mechanism aided the individual collective entities' role transitions. Training prior to system usage can help develop operator mental models regarding the system *usability* and how to interact with the recommended suggestion.

Providing suggestions to the operator about what actions can help transition individual collective entities into new roles promotes transparency by improving the system's *usability*. The transparency promoted in this design guideline needs to provide *explainable information* that allows the operator to *understand* what individual collective entities can transition into new roles and how the transition can be taken. *Control* mechanisms enabling the recommended suggestion must be provided in order to execute such actions. This level of transparency can alleviate *workload* by promoting better *SA*.

5.2.4.3 *Feedback* about Changing *Capabilities* or Roles

Relevant feedback can be provided to the operator in order to *justify* why individual collective entity capabilities or roles are changing, guideline DG_{29} in Table 5.7. Providing *feedback* promotes *explainability* by being *learnable*, which can improve operator *satisfaction* and *SA*, as well as calibrate operator *expectations* regarding how characteristics, such as environmental or system, influence changes in individual collective entity capabilities or roles. This guideline was formulated considering how operators often do not *understand* why collectives are behaving in a particular manner. Providing *feedback* is a useful method for promoting transparency; however, too much *feedback* may distract operators and can cause higher *workload*.

The design guideline can be embedded into the model, visualization, or both. The model can add identifiers to the *state information* provided from the collective, similar to DG_{22} in Chapter 5.2.7.2, that distinguish specific characteristics, such as perceptual

accuracy. A *process* that compares the identifiers across multiple time steps, in order to reduce noise associated with reported data, can determine whether a change is occurring in individual collective entity *capabilities* or roles. *Information* not provided over the allocated time steps due to limited communication, will likely defer to the last known *capability* or role *state information*, where providing an associated error or explanation of this occurrence may be needed. A threshold value derived using DG_{19} principles from Chapter 5.2.1.1, can be used to trigger sending a message to the visualization software that determines how to present the feedback to the operator.

The *feedback information* presented on the visualization must be *explainable* in order to *foster understanding* regarding why individual collective entity *capabilities* or roles are changing. The feedback *information* can be presented in various ways, including color coding, representative symbols, or icons that were discussed in Chapter 5.2.1.1. The characteristics contributing to changes in individual collective entity *capabilities* or roles can be presented as representative icons, such as a representation that a critical sensor is malfunctioning, in order to mitigate the amount of text provided to the operator. *Understanding* what is contributing to changes in *capabilities* or roles will help properly calibrate the operator's *expectations* of the collective's future behaviors and can improve *SA*. The *feedback* can be accessed as supplementary *information* in order to mitigate *workload* issues that may arise with too much supplied *information*.

Providing *feedback* to the operator will not necessarily *prompt* operators to proactively interact with the system; thus, *control* mechanisms that enable an operator to access the *feedback information* must be provided.

Providing *feedback* about what characteristics that contribute to *capability* or role changes promotes transparency by improving the system's *explainability*, which fosters the operator's *understanding*. This level of transparency can improve *SA* and provide

accurate *justifications* for collective behaviors.

5.2.4.4 Re-assigning Roles Strategies

Embedding strategies into the model *capable* of re-assigning individual collective entities into new roles if other members capabilities fall below a critical threshold is design guideline DG_{30} in Table 5.7. Providing the system *capability* to re-assign roles can be more *effective* and *efficient*, which *promotes* better system *usability*. This guideline was formulated considering how system *control* can help alleviate operator *workload* associated with determining information from Chapter 5.2.4.2, such as which individual collective entities can be transitioned into a new roles, how to make that transition, and when the transition must occur. Human-collective system designers must consider how to maximize the strengths of the model and operator in order to promote optimal *performance*. Using a model designed to achieve a task without operator influence will aid operators and can improve human-collective *performance*.

This design guideline is intended to be embedded into the model only. The principles outlined in Chapter 5.2.4.2 can be used to establish roles, distinguish individual collective entities' *capabilities*, and identify what roles the entities are currently performing. The model can monitor whether *capabilities* of individual collective entities have decreased below an acceptable threshold using the DG_{19} principles from Chapter 5.2.1.1. *Capabilities* that have surpassed the acceptable threshold can trigger a *process* to re-assign the lower *capability* individual collective entities' respective roles to other members who have better *capabilities*. Once the model has determined the best role re-assignment and method of executing the change it can be communicated to the visualization software that determines how to present the *information*.

Providing *information* on the visualization about the system's *capability* to re-assign roles is not necessary, rather operators can have access to the inputs, *information*, and decision-making conclusions the model outputs. Training prior to system usage is needed in order to promote accurate mental models of system *reliability* and *understanding* regarding system *processes*. Depending on the *reliability* or *process*, an operator may be prompted to influence the system by providing additional *information* to influence the system role re-assignment decision-making outcome or overrule the system's decision depending on the current situation. *Control* mechanisms to enable the operator to provide additional *information* or overrule the system's decision must be provided.

Embedding processes that enable models to re-assign individual collective roles promotes transparency by improving the system's *usability*. This level of transparency can alleviate operator *workload*, while promoting better human-collective *performance*. The model will initially have *control* over what individual collective entities are being re-assigned to different roles and how the re-assignment will be executed; however, operators will be *capable* of supplementing additional *information* to improve the re-assignment decision-making or override the model.

5.2.5 Limited Communication Among Individual Collective Entities

The biologically inspired behavior of focus is that the communication amongst individual collective entities is typically limited. A topological method of communication [181] describes the behavior of biological collectives that communicate with a particular number of neighbors, which varies based on species, such as six to seven for starlings [6] and

three to fix for shoaling fish [21]. Differences in the number of neighbors individuals are capable of communicating with vary depending on particular characteristics. Robotic collectives will experience visual communication [181] challenges similar to those seen in fish, such as some individuals may visually occlude others' *capability* to *observe* neighbors [22] due to their size or the density of the aggregation. Interacting with few neighbors can reduce noisy *information*; however, the *information* is short ranged [6]. Reshuffling individuals in the group changes exposure of neighbors and provides different *information*, which is a biological fish strategy [6] that can be embedded into robotic systems. Much of the existing literature, including the single human-collective evaluations from Chapter 3, have assumed perfect communication between the individual collective entities and between the operator and the collectives. Designers of collective systems must consider how limited communication amongst the individual collective entities and between the operator and a collective can influences human-collective interactions and *performance* in order to operate in real-world use scenarios.

Table 5.8: Design guidance for limited communication amongst individual collective entities.

DG_{31}. Indicate collective communication *status* to the operator, such as high-bandwidth level.
DG_{32}. Provide *feedback* to the operator about communication *status* implications, such as limited-bandwidth level means limited communication with the respective collective and delays in behavior response are expected.
DG_{33}. Provide *predictive information* about delay in updates related to collective behavior due to the communication *status* and the error associated with the *prediction* to the operator.

The design guidance is to train operators about limited communication *capabilities* associated with the system prior to usage, provide communication *state information*, and to provide *feedback* about the communication *status* implications on collective behavior

response. *Predictive information* about the delay in *information* updates and *information* about communication latency are also recommended. Operators who have a better *understanding* of communication limitations will have more accurate *expectations* of the system *capabilities* and *performance*, which will improve operator *satisfaction*.

5.2.5.1 Communication *Status*

Communication bandwidth or locations at which communication is unavailable are impacted by the operational environment and can negatively influence the ability of the operator and collectives to *effectively* communicate and interact with one another, which may cause poor human-collective *performance*. Indicating what the collectives' current communication *status* is to the operator is design guideline DG_{31} in Table 5.8. This guideline was formulated considering how challenging it may be for operators to *understand* what is influencing collective behaviors, such as system or environmental characteristics. Providing the current communication *status* of the collective majority, since variability will exist among individual entities, will improve the operator's *SA*, calibrate the operator's *expectations* accurately, and reduce the *workload* associated with determining whether the *information* provided to and from the collective was received. System *usability* will improve by indicating the collective's current communication *status*, because operators will have a better *understanding* of the communication *reliability*.

The design guideline can be embedded into the model, visualization, or both. The model can aggregate at minimum a particular percentage of available *information* provided from the collective, since variability or lack of communication may impact the ability for individual collective entities to communicate. A current communication *status* can be computed, such as bandwidth availability, latency, or signal strength. After

calculating the communication *status*, the model can send a message to the visualization software that determines how to present the *information* to the operator.

The *information* presented on the visualization must be *observable* in order to *foster understanding*. Providing a clear and *legible* communication *status* message can be embedded into the visualization using various techniques. A succinct text message, such as "Low", lower opacity color, or a vertical bar indicator at a low level can represent low bandwidth, while "High", an opaque color, or a high level on a vertical bar indicator can represent high bandwidth. Using a clear and succinct text message can be advantageous, as long as the operator's *workload* is not negatively impacted by adding more text to the visualization. Color usage can expedite detection; however, particular colors must be avoided in order to accommodate operators who may be color blind or to deconflict the chosen color with the use of color on the visualization for other purposes. The levels on a vertical bar indicator must be *observable* and distinct, if not operators may interpret the same level differently, which may negatively impact *SA* and misalign operator *expectations* of the system *usability*. The presentation of the communication *status* can be accessed in supplemental windows associated with the collectives in order to reduce visualization clutter.

Providing the communication *status* of the collectives may impact the operator's interactions with the system, as long as training prior to system usage developed accurate operator mental models of the implications of the communication *status*, such as bandwidth level, latency, or signal strength. The communication *status* will either affirm the operator's *understanding* of the ideal system *capabilities* and not influence their interactions with the system, or will *prompt* the operator to re-calibrate their *expectations* based on the communication *reliability*. A lower communication *status* may cause operators to limit their interactions with the system until the communication *status* increases to an

optimal or desirable communication level or strength. *Control* mechanisms that allow operators to access supplemental communication *status information* must be provided.

Providing the collectives' communication *status* promotes transparency by improving the system's *usability* and *explainability*. The transparency promoted in this design guideline needs to provide *observable information* in order for the operator to *understand* the collectives' communication *status*. This level of transparency can promote better *SA* and accurately calibration operators' *expectations* of system response. *Understanding* the *reliability* of the communication can improve human-collective interactions.

5.2.5.2 *Feedback* about Communication *Status* Implications

Relevant feedback can provide *context* and insight regarding the implications of communication *status* on human-collective interactions and behaviors, design guideline DG_{32} in Table 5.8. Providing *feedback* to the operator promotes *explainability* by being *learnable*, which can improve operator *satisfaction* and *SA*, as well as calibrate operator *expectations* regarding the system's *reliability*. This guideline was formulated considering how operators often do not *understand* the implications of changes in the system, such as communication availability, and how that influences the operators' ability to interaction with the collectives. Providing *feedback* is a useful method for promoting transparency; however, too much *feedback* may distract operators and can cause higher *workload*.

The design guideline can be embedded into the model, visualization, or both. The model can aggregate at minimum a particular percentage of available *information* provided from the collective, due to variability or lack of communication. The model can compute a communication *status*, such as the example provided in DG_{31} in Chapter 5.2.5.1. The variables used to compute the communication *status* can indicate whether

particular *information* is missing, due to signal detection issues for example, or if the available *information* is causing substandard communication. Situations where no information is provided, because there is no communication, the system can report the last known communication status and indicate to the operator that the information is not current, rather from a prior time. A set of predetermined general implications on human-collective interactions can be encoded into the model that is dependent on the communication level computed. The model can send a message to the visualization software, after determining the communication implications on the human-collective team, that determines how to present the *feedback* to the operator.

The *feedback information* presented on the visualization must be *explainable* in order to *foster understanding* regarding what the communication *status* is, what factors (i.e., variables) are influencing the level of communication, which is only needed when the communication is below an optimal level, and what are the implications on human-collective interactions and behaviors. The communication status can use the text, color, or icon strategies discussed in DG_{31} in Chapter 5.2.5.1. Explanations about what factors are influencing communication and what the implications are on human-collective interactions and behaviors can be provided via text; however, representative symbols can also be used. Presenting *information* using particular techniques will be dependent on the level of detail needed to provide sufficient *context*, whether legends are available to remind operators what symbols represent, and if prior training can aid operator *understanding* of the *information* provided on the visualization. Designers must balance how *information* is presented in order to mitigate *workload* associated with identifying and *understanding* the *feedback* as well as not exceeding operator cognitive *capabilities* by requiring recollection of the meaning of colors, symbols, or icons. Providing *information* regarding the communication implications on human-collective interactions and be-

haviors will help calibrate properly the operator's *expectations* of the collective's future behaviors, can improve *SA*, and provide insight about system *capabilities*. The *feedback* can be accessed in a supplementary *information* window in order to mitigate *workload* issues associated with providing too much *information*.

Providing the communication implications on human-collective interactions and behaviors will ideally *prompt* users to either continue (high bandwidth) or lessen (lower than high bandwidth) the number of interactions with the system. Fewer interactions with the human-collective system may persist until the communication level has returned to a more optimal level, which may positively or negatively influence human-collective interactions and behaviors. *Control* mechanisms must be provided to the operator in order to access the communication implications *feedback* and influence the collectives accordingly. Designers must consider how influential *control* mechanisms must be in order to accommodate situations at various communication levels.

Providing *feedback* regarding the implications of communication on human-collective interactions and behaviors promotes transparency by improving the system's *explainability* and *usability*. The transparency promoted in this design guideline needs to provide *explainable information* that *fosters* operator *understanding* regarding why factors are influencing communication *status* and how communication *status* will impact the *effectiveness* of human-collective interactions and behaviors. This level of transparency can improve *SA*, lower *workload*, and provide accurate *justifications* for operator actions.

5.2.5.3 Delay in Updates *Prediction* and Error

Providing an operator *predictive information* related to communication *status* implications, such as delays, on human-collective interactions and behaviors, as well as the

associated *prediction* error is design guideline DG_{33} in Table 5.8. This guideline was formulated in order to provide operators an accurate *understanding* of when they can interact with the system due to communication delays rather than the time needed for collective behavior to stabilize (e.g., neglect benevolence). Providing *relevant predictive information* can improve *efficiency*, *effectiveness*, and *timing* associated with waiting for the communication to return to a level that permits desired human-collective interactions and behaviors. The operator; however, will need *information* related to how accurate the *prediction* is and what is the system's confidence in the provided *prediction*, similar to the example provided in the DG_{19} example in Chapter 5.2.1.1. The *prediction* error will inform the operator about the system's confidence (i.e., *reliability*) in the *prediction*, as well as calibrate the operators' *expectation* and *reliance* on the system, which promotes better system *usability*.

The design guideline can be embedded into the model, visualization, or both. The model can aggregate the *information* from the collective and compute a communication *status* (DG_{31} in Chapter 5.2.5.1). Providing *predictive* delay *information* is needed when the communication *status* is below an optimal level causing delays. Situations with suboptimal communication status' can use a decision support tool or *predictive* simulation tool from DG_{21} in Chapter 5.2.1.3, in order to calculate the probability of a collectives' communication *state* changing and predict how much *time* it will take to reach an optimal *status* where the operator has the *capability* to influence the collective and the resulting behavior emerges. The system can send a message to the visualization software that determines how to present the communication delay information to the operator.

The *information* presented on the visualization must be *observable*, *explainable*, and *understandable* by the operator. Providing a clear and *legible* message indicating the *predictive information* related to communication *status* implications can be embedded into

the visualization using a representative icon. A stopwatch symbol, for example, can be used to draw the operator's attention to the icon that shows the time delay, for example, using numerical values and the associated units. A stopwatch symbol may not always indicate the *predictive information*, since the characteristics used to calculate communication *status* and make *predictions* may change. Choosing a consistent symbol, regardless of the *predictive information* and communication *status* characteristic, is recommended in order to maintain consistent mental models. The *predictive* error can also be provided on the icon using numerical values and a percentage sign in order to ensure distinction between the two numerical values. Representative symbols or icons are an *effective* alternative that minimizes the use of text and can be perceived quickly. The *predictive* delay and error *information* can be provided on a supplemental pop-up window near the collective in order to mitigate the *workload* associated with the amount of *information* provided on the visualization. The supplemental window can remain visible for operators only when the communication status is suboptimal.

Providing *predictive information* related to communication *status* implications and the associated *prediction* error may *prompt* users to continue interacting with the system only when the communication *status* has returned to a more optimal level. Designers must consider how providing the *predictive information* may discourage operators to interact with the system, which may cause poor human-collective *performance*. Model strategies, such as *predicting* what happens when the operator takes particular actions, can help mitigate disuse when the communication *status* does not return to an optimal level. Operators must *understand* the implications of communication *status*, DG_{32} in Chapter 5.2.5.1, and what interactions can still be taken during delays. Training prior to system usage can ensure more accurate mental models of system usage during limited communication situations. *Control* mechanisms must be provided to the operator in

order to access the *predictive* delay *information* and associated error, as well as enable the operator to influence the collectives accordingly.

Providing *predictive* delay *information* and the associated error promotes transparency by improving the system's *usability*. The transparency promoted in this design guideline needs to provide useful and *explainable information* that allows the operator to *understand* the collective's future *state* and what interactions the human-collective team can take during situations with limited communication. This level of transparency can promote better *SA* and mitigate *workload*.

5.2.6 Collective and Subgroup *Information*

Information provided from the collective can be presented at various aggregated levels, including the collective (high aggregation) and subgroup levels (low aggregation). Providing *information* at the individual collective entity level (no aggregation) is not recommended due to the quantity of *information* provided by very large sized collectives, the noise of the provided *information*, and the limited computational *capabilities* of the system to *process* the *information*. Local sampling, performed in parallel by large numbers of individuals, allows biological colonies to accurately average individual members' responses to changes, such as environmental changes [14]. The group level reporting of *information* mitigates the noisy individual level reporting of *information* [5]. Situations may arise when operators want to access collective and subgroup level *information*. Providing the operator access to different *information* levels can improve their *understanding* of how subgroup behaviors influence collective behaviors and visa versa, as long as the *information* is presented clearly and distinctly. Designers must determine how many subgroups can be presented to the operator before exceeding the operator's cognitive

capabilities, if many subgroups exist within the collective.

Table 5.9: Design guidance for presenting information about the collective and subgroups.

DG_{34}. Provide *feedback*, with an associated error, to the operator about why a collective or subgroup is doing what it is currently doing, such as fulfilling a particular task or reacting to an environmental perturbation.
DG_{35}. Provide *predictive information*, as well as the error associated with the *prediction*, to the operator about how the behavior of the collective or subgroup may influence the overall collective's *state* and actions.
DG_{36}. Provide suggestions to the operator about how to mitigate or support collective or subgroup behavior.
DG_{37}. Only provide *information* to the operator about subgroups if collective *state* will change significantly (e.g., critical threshold reached).

The design guidance is to provide feedback to the operators about what particular behavior or action a collective or subgroup is exhibiting or performing and why that behavior or action is occurring. Providing predictive collective or subgroup information and suggestions about how the operator can mitigate or help support particular behaviors is also recommended. Information about subgroups is recommended only when a significant collective state change will occur. Operators who have a better understanding of the current and predicted behaviors of the collective and subgroups will be able to interact with the collective and subgroups more effectively, which will ultimately improve task performance.

5.2.6.1 *Feedback* about Collective and Subgroup Actions

Relevant feedback and the error (e.g., confidence level) associated with the *feedback* can be provided to the operator regarding why a collective or subgroup is doing what it is currently doing and what may be contributing to that particular behavior, guideline

DG_{34} in Table 5.9. Providing *feedback* promotes *explainability*, which can improve operator *satisfaction* and *SA*, while providing the associated error calibrates operator *expectations* and provides insight about the system *reliability*. This guideline was formulated considering how operators often do not *understand* why collectives are behaving in a particular manner and what is contributing to those behaviors. Providing *feedback* is a useful method for promoting transparency; however, too much *feedback* may distract operators and can cause higher *workload*. There will be a trade off between the quantity and quality of *feedback* provided to operators. Systems composed of many collectives and their respective subgroups must have a limited number of messages provided to the operators, as well as fewer details provided in the *feedback*, in order to mitigate *workload*, contributing to confusion, and potentially hindering human-collective *performance*.

The design guideline can be embedded into the model, visualization, or both. Prior to *process*ing the data provided from the collective, a set of general characteristic classifiers can be embedded into the model that estimate how many individual collective entities are in particular subgroups based on their *capabilities*, roles, and other useful characteristics associated with the human-collective task. The collective and subgroup behavior *states* can be determined in order to *understand* what the collective or subgroups are doing. Characteristics, such as environmental perturbations, must be determined in order to provide the operator with an explanation of what is contributing to the collective or subgroup behavior *state*. All of the available *information* can be aggregated into the collective level reporting, and categorized by a general characteristic classifier for the subgroup level. Different criteria can be used to determine what subgroups can be visually presented to the operator. Aspects, such as the quantity of individual collective entities sharing the same general characteristic classifier, the quantity of influence particular subgroups possess, or operator selected subgroups, can be used to identify

which subgroups will be presented to the operators. A decision support tool or *predictive* simulation tool from DG_{21} in Chapter 5.2.1.3 can be used to calculate the quantity of influence subgroups have on collective task *performance*, as well as the associated error, such as a confidence interval. The quantity of subgroup information presented to the operator must be limited, such as seven plus or minus two, which is consistent with limited human short term memory capacity [180]. The confidence interval can be used to trigger sending a message to the visualization software that determines how to present the *feedback* to the operator.

The *feedback information* presented on the visualization must be *explainable* in order to *foster understanding* regarding what the collective and its subgroups are doing and why, as well as the associated feedback error. The *feedback information* can be presented in various ways, including color coding, text messages, representative symbols, or icons that were discussed in Chapter 5.2.1.1. A designated interactive feedback area can be used to present the *information* and can remain visible throughout the duration of system usage. The interactive area can be subdivided into four sections. The operator can 1) select a respective collective in order to 2) identify what subgroups are performing specific tasks. The operator can select either the collective or a respective subgroup, which will be highlighted upon selection, in order to see what 3) the collective or respective subgroup is doing. The last section can 4) identify what may be contributing to those collective or subgroup actions and the associated error. Using a static designated area on a visualization can be advantageous if a potentially large quantity of *information* can be provided from the system. The designated area can show the hierarchy of collective *information* (collective as top level and subgroup as a lower level) so that operators *understand* what behavior they are *perceiving*. A consistent location and presentation strategy will aid operator mental models of system *usability* and *explainability*. Designated

interactive areas may not be feasible for all visualizations if the visualization does not have sufficient space for an additional designated area and if adding the area does not contribute to clutter. Designers must consider how to balance *information* presentation by using various color, pattern, text, static persistent *information* presentation, or supplemental *information* presentation strategies in order to mitigate workload associated with perceiving and comprehending the *information*. *Understanding* what the collective or subgroup is doing and what may be contributing to those actions will help calibrate the operator's *expectations* of the collective's future behaviors properly, improve *SA*, and provide insight about system *capabilities*.

Providing *feedback* to the operator will not necessarily *prompt* operators to proactively interact with the system; however, *information* presentation techniques, such as the designated interactive area, may motivate operators to inquire *feedback* about current collective or subgroup actions. A well designed designated interactive area must have *control* mechanisms that enable an operator to access desired *feedback information*. An overly complex *feedback* presentation may demotivate operators to use the *information*. Designers must balance between the quantity and quality of *information* provided to the operator, as well as what control mechanisms enable operators to *effectively* interact with the system. Training prior to system usage can also aid operators to develop accurate mental models of system *usability* and *explainability*.

Providing *feedback* about what a collective or its subgroups are doing, what may be contributing to those actions, and the associated error of the *feedback* promotes transparency by improving the system's *explainability*, which fosters the operator's *understanding*. This level of transparency can improve *SA* and provide accurate *justifications* of collective or subgroup actions. Designers must be cognizant about how to provide the *feedback* in order to mitigate negative influence on *workload* with perceiving and

comprehending the *information*.

5.2.6.2 *Prediction* and Error of Collective or Subgroup Influence

Providing an operator a likelihood *prediction* of how a collective or subgroup's current actions may influence the collective's future *state* and actions is design guideline DG_{35} in Table 5.9. This guideline was formulated considering how operators often do not *understand* what the implications of current collective or subgroup actions are on the future collective *state*. Providing *relevant predictive information* that can inform the operator about future collective *states* can mitigate *workload*, as well as help improve *SA* and human-collective task *performance*. The operator will need *information* related to how accurate the *prediction* is and what is the system's confidence (i.e., *reliability*) in the provided *prediction*, such as a confidence interval, in order to have accurate *expectations* and *reliance* on the system, which will promote better system *usability*.

The design guideline can be embedded into the model, visualization, or both. Strategies that can process the available collective *information*, classify the *information* by collective or subgroups, and determine what the collective and respective subgroups are doing were discussed in design guideline DG_{34} in Chapter 5.2.6.1. Based on the influence of the current collective's and subgroups' actions, a *prediction* and an associated *prediction* error can be computed using a *predictive* simulation tool from DG_{21} in Chapter 5.2.1.3 in order to estimate the collective's future *state*. A message to the visualization software can be sent that determines how to present the *predictive information* an associated *prediction* error to the operator.

The *information* presented via the visualization must be *observable*, *explainable*, and *understandable* by the operator. Providing a clear, succinct, and *legible* explanation in-

dicating what will be the collective's future *predicted state*, what collective or subgroup actions are contributing to that *prediction*, and the *prediction* error can be embedded into the visualization. The collective's *predicted* future *state* and *prediction* error can be presented on the collective icon using a letter to represent the most supported state next to the associated error, such as F (Error 12%), where F represents favoring a particular target. The collective *predictive state*, *prediction* error, and details regarding what collective or subgroup actions are contributing to the *predicted state* can become visible on the visualization when the operator requests to view the *information*. Specific details regarding what collective or subgroup actions are contributing to the *prediction* can be provided via a supplementary window. A legend providing *information* about what the collective future *state* acronyms represent can alleviate *workload* associated with remembering the representative *information* and ultimately improve *SA*. *Understanding* the collective *predictive state* and error due to current collective or subgroup actions will indicate how *reliable* the system is and will properly calibrate the operator's *expectations* of the collective's future behaviors.

Providing the collectives' *predicted* future *state* along with a *prediction* error will ideally *prompt* operators to proactively interact with the system. The operator will likely want to redirect the collective's or subgroups' current actions in order to produce a desired future collective *state*. *Control* mechanisms must be provided in order to enable the operator's attempt to redirect collective or subgroup behaviors, as well as access the supplemental windows detailing what actions are contributing to the *prediction*. Training prior to system usage can help operators develop accurate mental models of system *usability* in order to expedite *information* processing during system usage.

Providing a *prediction* and an associated error of a collectives' future *state*, as well as an explanation of what current collective or subgroup actions are contributing to that

prediction promotes transparency by improving the system's *usability* and *explainability*. This level of transparency can promote better *SA*, which prompts the operator to take appropriate preventative actions that result in improved human-collective *performance*.

5.2.6.3 Suggestions to Support or Mitigate Collective or Subgroup

Providing suggestions to the operator about what actions can be used to mitigate or support collective or subgroup behavior is design guideline DG_{36} in Table 5.9. Providing the operator with suggestions can alleviate the *workload* associated with determining how to prevent, minimize, support, or increase further development of collective or subgroup behaviors. This guideline was formulated considering how the system can proactively aid the operator in the development of particular collective or subgroup behaviors. Providing *relevant* suggestions can help improve *efficiency*, *timing*, *effectiveness*, and the overall human-collective's task *performance*. Human-collective system designers must consider how providing system suggestions influences operator behaviors and *expectations*. Some operators may become overly *reliant* on the system and expect immediate behavior development, although time is needed for the collective behavior to stabilize (i.e., neglect benevolence). Deviations from these operator *expectations* can cause dis*satisfaction*, reduce system *credibility*, and perceived *reliability*. The operators may overcompensate for a lack of collective or subgroup behavior development by taking *control* of all proceeding actions and ignoring what the system is reporting, which can negatively impact system *usability*. Providing additional *information*, such as *feedback* about the *time* remaining for the collective or subgroup to change behaviors, and training prior to system usage can help mitigate operator *expectation* misalignment issues and improve operator interactions with the system.

The design guideline can be embedded into the model, visualization, or both. The model can initially estimate the collectives' or subgroups' current *state* from the *information* provided by the collective, after classifying what information is associated with particular collectives and respective subgroups using principles from DG_{34} in Chapter 5.2.6.1. A deviation from the expected or desired *state*, similar to that discussed in DG_{20} from Chapter 5.2.1.2, can be calculated. A threshold value derived using the DG_{19} principles from Chapter 5.2.1.1, can be used to trigger a decision support tool or predictive simulation tool, which can calculate what actions will contribute to a lower deviation from the expect or desired *state*. Once the model has determined the best suggestion, the *information* can be communicated to the visualization software that determines how to present the *information* to the operator.

The *information* presented on the visualization must be *observable* and *explainable* in order to *foster understanding* of the longer term implications of the potential suggestions. Providing a clear, succinct, and *legible* suggestion indicating what action the system recommends and why for a particular collective or subgroup can be embedded into the visualization. The presentation of the *information* can be conveyed on a pop-up window near the collective. A text message can indicate what particular *control* mechanism, for example "investigate" target, is suggested and identify whether the suggestion is for the collective or a particular subgroup. The text must be easily *observable* (e.g., large easily visible letters) and *understandable* in order to be an *effective* implementation. The pop-up window can remain visible to the operator until either the suggestion is accepted, canceled, or the *time* to issue that particular suggestion expires.

Providing suggestions about what operator actions can mitigate or support further development of collective or subgroup behavior will ideally *prompt* operators to interact with the system. *Control* mechanisms must be provided that enable the operator's at-

tempt to mitigate or support behavior for a collective or a subgroup. The operator must also have the *capability* to adjust software parameters used to determine suggestions, as adjustments to the parameters can result in different suggestions to mitigate or support collective or subgroup behavior. Recording which *control* mechanisms have been used can help promote operator *understanding* of whether or not using the particular *control* mechanism mitigated or supported the collective or subgroup behavior. Training prior to system usage can develop accurate operator mental models regarding the system's *usability* and how the operator can interact with the system in order to fulfill, ignore, or modify the recommended suggestion.

Providing suggestions to the operator about what actions can mitigate or support collective or subgroup behavior development promotes transparency by improving the system's *usability*, *explainability*, and human-collective *performance*. The transparency promoted in this design guideline needs to provide *information* that allows the operator to *understand* what actions can mitigate or support further development of collective or subgroup behavior. *Control* mechanisms must enable operators to execute such actions to the particular collective or subgroup. This level of transparency can alleviate *workload* by promoting better *SA*.

5.2.6.4 Presentation of Subgroup *Information*

Providing *information* to an operator requires determining what types of *information* are necessary, how best to present the *information*, and how often the *information* must be made available to the operator. Only providing subgroup level *information* to the operator, if the collective *state* will change significantly as a result of the subgroup behavior, is DG_{37} in Table 5.9. Providing *information* when an operator needs it, which was the

inspiration to formulating this design guideline, will promote better *usability*, because the system will be more *effective*, *efficient* and *timely*. Operators may experience less *workload*, since only *relevant information* is provided when it is needed. Operators will be *reliant* on the system to provide the *information* when they need it; therefore, their *expectation* is that the system is *capable* of fulfilling that objective. Any complications, such as limited communication, that may hinder the system's *capability* of supplying subgroup *information* when the operator thinks the *information* must be presented will negatively influence *credibility* and perceived *reliability* of the system.

This design guideline is intended to be embedded into the model only. The model can implement the strategies proposed in design guideline DG_{34} in Chapter 5.2.6.1 that process the available collective *information* and classify the *information* by collective or subgroups. A threshold value derived using DG_{19} principles from Chapter 5.2.1.1, can be used to trigger sending a message to the visualization software that determines how to present the *information* to the operator.

Providing subgroup level *information*, such as feedback, current and predictive state information, and suggestions, on the visualization were discussed in Chapters 5.2.6.1, 5.2.6.2, and 5.2.6.3 as well as the respective *control* mechanism implications. The operator must have the *capability* to adjust software parameters used to determine which subgroup behavior is presented to the operator, as adjustments to the parameters can result in different outcomes.

5.2.7 Leadership

Many homogeneous robotic collective systems have been used to assess bio-inspired behaviors; however, biological collectives are typically heterogeneous due to various

characteristics. One type of characteristic that can make systems heterogeneous is leadership, which can be crucial for biological collectives survival. The survival of a honeybee colony, for example, is dependent on the survival of its queen, who carries the new colony's genes [5], and which makes her a leader in this particular *context*. Individual collective entities can become leaders based on a particular context, how much information they possess, environmental effects, their experience, and possessing certain physiological (e.g., software or hardware) characteristics. Scout honeybees can be considered leaders, since their knowledge informs when to initiate the departure of the daughter colony from the mother colony, how to chose and make a decision on a suitable nest, trigger the colony's takeoff to the new nest site, and steer the colony during its flight [5]. Environmental aspects, such as the distance between locations and predictable resources, can cause leadership to become transient, as observed in dolphins [26]. *Understanding* what characteristics can determine leadership and how that leadership can change are necessary in order to design *effective* heterogeneous human-collective systems. Providing the operator *information* regarding the leaders' *state*, influence level, and the *reliability* of the influence, will aid the development of accurate mental models of leadership *usability* and promote *SA*.

Table 5.10: Design guidance for leadership.

DG_{38}. Provide *information* about the leaders' *state* in relation to a goal and influence on the collective, such as the rate of change in behavior.
DG_{39}. Provide *feedback* to the operator about a leader's or a group of leaders' influence level, such as *time* to complete tasks, and the reliability (i.e., can the leaders influence the entities how the operator intends).

The design guidance is to provide the leaders' status with respect to the mission goal and feedback about the leaders' influence on the collective. Operators who have

a better understanding of how to use the leaders' to influence collective behaviors will have more effective interactions and improve the human-collective team's performance.

5.2.7.1 Leader *State* and Influence on Collective

Providing information regarding leaders' *states* in relation to a goal and the influence those leaders have on a collective is design guideline DG_{38} in Table 5.10. This guideline was formulated considering how challenging it may be for operators to *observe* and *understand* leaders' states, especially when using abstract visualizations that do not show individual leaders to the operator, and what the leaders' level of influence is on the collective and its overall mission goal. Providing the leaders' current *state* will attract the operator's attention to *relevant information*, which can improve their *SA* and reduce the *workload* associated with determining what is happening with the leaders. *Relevant information* about the leaders' persistent and transient *capabilities* (e.g., influence on the collective) will promote *learnability*, as operators will be able to use leadership more *effectively* to influence collective behaviors to aid in achieving a desired mission goal.

The design guideline can be embedded into the model, visualization, or both. Prior to *process*ing the data the model can use principles from DG_{34} in Chapter 5.2.6.1 to classify individual collective entities versus leaders based on general characteristics, such as their *capabilities*, roles, or operator selection. The model can aggregate the available *information* from the collective, some of which may be missing due to imperfect communication, and compute a current collective and leaders' *state*, with respect to the given task and environment *state*. A *process* will need to be embedded into the model in order to determine the leaders' influence on the collective with respect to the mission goal, for example, calculating the rate of change in collective behavior after being influenced by

particular leaders. Criteria discussed in design guideline DG_{34} in Chapter 5.2.6.1 can be used to determine what leaders' *state information* and influence on the collective can be visually presented to the operator. The quantity of leader information presented to the operator must be limited, such as seven plus or minus two, in order to avoid overloading limited human short term memory *capacity* [180]). A threshold value derived using DG_{19} principles from Chapter 5.2.1.1, can trigger sending a message to the visualization software that determines how to present the *information* to the operator.

The *information* presented on the visualization to the operator must be *observable* and *explainable* in order to *foster understanding*. Providing a clear and *legible state* message regarding the leaders' *state* and influence on the collective can be embedded into the visualization. The presentation of the *information* can be conveyed using an aggregated representative icon, such as a glyph [40], in order to reduce the quantity of associated text that may cause mis*understanding* and confusion. A glyph is a simple visual icon that depicts multiple attributes [182]. A collective hub icon, for example, can be treated similar to a glyph with added *information* about the leaders' *state* and influence on the respective collective. Designers can use different colors or patterns to indicate the leaders' state, while a dedicated area consumed by a particular leader can indicate its influence on the collective. Representative symbols can lower *workload* by being easier to *observe* quickly and to *understand*. Designers must balance the amount of text versus symbols or icons in order to alleviate workload. The *usability* of glyphs may become disadvantageous as the number of collectives represented on a visualization increases, as well as the complexity or quantity of the data represented on the glyphs.

Providing the leaders' *state information* and their influence on the collective will ideally *prompt* operators to interact with the system in order to ensure mission completion. *Control* mechanisms must be provided to enable the operator's attempt to help

direct leaders' in order to positively influence the collective towards mission completion. Recording which *control* mechanisms have been used is necessary for the operator to *understand* whether using the *control* mechanism promoted positive task completion. Training prior to system usage can help operators develop accurate mental models of how particular *control* mechanisms direct leaders' to influence collectives positively.

Providing the current leaders' *state* and the influence the leaders' have on the collective with respect to the mission goal promotes transparency by improving the system's *usability* and *explainability*. The transparency promoted in this design guideline needs to provide *observable* and *explainable information* to *understand* the leaders' current state and influence on the collective. This level of transparency can promote lower workload, by reducing the amount of processing required to determine the influence leaders have on a collective, and can improve *SA*, which prompts the operator to take appropriate preventative actions that result in improved human-collective *performance*.

5.2.7.2 Leadership Influence and Reliability *Feedback*

Relevant feedback can provide the operator *information* regarding the influence level of a leader or group of leaders' and how reliable that associated influence may be, guideline DG_{39} in Table 5.10. Providing *feedback* promotes *explainability* of system limitation *capabilities*, which will impact how the operator interacts with the system. Operator *satisfaction* and *SA* can increase by knowing *relevant* leadership *information* that can improve the *effectiveness* of operator *directability*. This guideline was formulated considering how operators may not *understand* the *reliability* and level of influence leaders have over the collective. Providing *feedback* is a useful method for promoting transparency; however, too much *feedback* may distract operators and can cause higher *workload*.

The design guideline can be embedded into the model, visualization, or both. The model can embed principles discussed from DG_{38} in Chapter 5.2.7.1 regarding classification of individual collective entities versus leaders based on characteristics, such as their *capabilities*. The model can use a decision support tool or predictive simulation tool, similar to that from DG_{21} in Table 5.4, that can *predict* how the leader classification influences the collective behavior relative to mission completion. An associated error with the influence *prediction* can be calculated. Criteria discussed in design guideline DG_{34} in Chapter 5.2.6.1 can be used to determine what leaders' influence *information* can be presented visually to the operator. The quantity of leader information presented to the operator must be limited in order to avoid overloading human short term memory *capacity* [180]). A threshold value derived using DG_{19} principles from Chapter 5.2.1.1, can trigger sending a message to the visualization software that determines how to present the *information* to the operator.

The *feedback information* presented on the visualization must be *explainable* in order to *foster understanding* of a leader's or groups of leaders' influence on the collective and associated error. The *feedback information* can be presented in various ways, including color coding, text messages, representative symbols, or icons that were discussed in Chapter 5.2.1.1. The presentation of the *information* can be conveyed using the glyph design mentioned from DG_{38} in Chapter 5.2.7.1. The influence prediction error can be represented using different colors or patterns, while a dedicated area consumed by a particular leader or group of leaders can indicate the influence level on the collective. Similar advantages and disadvantages associated with glyphs from DG_{38} in Chapter 5.2.7.1 are applicable for this design guideline.

Providing *feedback* to the operator regarding the influence leaders or groups of leaders have on a collective and the reliability of that influence will likely prompt operators

to interact with the system if influence and reliability are high, if not operators may decide to use other influence strategies. *Control* mechanisms must be provided in order to enable the operator's attempt to help direct leaders' to positively influence the collective towards mission completion. Recording which *control* mechanisms have been used is necessary in order for the operator to *understand* whether using the *control* mechanism promoted positive task completion. Training prior to system usage can help operators develop accurate mental models of how particular *control* mechanisms direct leaders' to influence collectives positively.

Providing *feedback* to the operator about the leader's or group of leaders' influence and reliability of that influence on collective behaviors promotes transparency by improving the system's *explainability*, which fosters the operator's *understanding*, as well as system *usability*. This level of transparency can promote lower workload, by reducing the amount of processing required to determine the influence leaders have on a collective, can improve *SA*, and provides *justification* to take particular actions that result in improved human-collective *performance*.

5.3 Design Guidance Reliability for Real World Use Scenarios

Limitations associated with real world use scenarios were briefly identified in some of provided design guidelines. Further discussions will expand on each limitation category: 1) limited or no communication, 2) challenges with the domain (e.g., aerial or underwater) or environment, and 3) the type of collective systems. This discussion is intended to aid designers' understanding of the limitations and how they may impact the reliability of the design guidelines for their respective human-collective systems performing specific tasks in particular environments. Understanding the limitations of

the associated design guidelines will aid future human-collective system evaluations, which will provide data to further validate the guidelines' reliability.

Limited or no communication situations may arise due to perceptual issues experienced in biological species, which was discussed in Chapter 5.2.5, as well as environmental, hardware, or software issues. Guidelines suggesting to provide particular types of information to the operator will be ineffective if the operator and collective system have limited to no communicate with one another, as will be guidelines with respect to individual collective entities communicating with collective hubs or with one another. Different types of queuing strategies may need to be considered in order to prioritize what information is most important or necessary in order to progress towards goal completion when communication is limited, and to determine how operator issued commands and system processes will be managed in order to avoid undesired behaviors, such as latency issues or detrimental collective behaviors.

Environmental characteristics, such as humidity, atmospheric pressure, temperature, and physical barriers (e.g., objects in the environment or building structures), may cause signal interference and can hinder the reliability of the hardware or software. Determining an acceptable range of operational conditions prior to system use are necessary in order to program appropriate mitigation strategies into the models and to determine bounding conditions for particular types of individual collective entities, such as operational altitude range. Many of the environmental characteristics will be challenging to determine a priori due to their variability and the inability to identify all possible environmental characteristics, such as changing turbidity of ocean environments. The working environment will also affect bandwidth due to the distance between the individual collective entities, collective hubs, and operators, as well as the communication system coverage patterns, which will ultimately affect the communication status of the

human-collective system.

Determining what types of hardware and software are needed for the overall encompassing system architecture, collective hubs, and individual collective entities, is necessary in order to ensure operation across mission domains and deployment environments. The use of different types of vehicles, such as multirotor versus fixed wing aerial vehicles, or surface versus underwater marine vehicles, with differing capabilities and functionalities, as well as differing domain characteristics will impact the reliability and generalizability of the transparency design guidelines. Complexity of a system will increase as collectives change from homogeneous to heterogeneous systems and will require guidelines compensating for varying individual collective entity variability, such as the guidance related to leadership. As collective systems become more sophisticated, such as improvements in perception, autonomy, and intelligence, so will the hardware and software required to implement those capabilities. Providing more design guidelines and system capabilities will contribute to increased system complexity and must be considered prior to implementation. The size of the collective may restrict hardware and software advances if the desired capability changes are too costly. Improved individual collective entity capabilities can assist with the implementation of the transparency design guidelines; however, the cost of more sophisticated individual collective entities will impact the overall cost of acquiring and maintaining the collective. The acquisition costs alone of a large collective (e.g., 1000 entities), where each entity has somewhat sophisticated capabilities (e.g., $1,500 per entity) will fundamentally cost more (e.g., $1,500,000), when compared to less capable entities (e.g., $150 per entity) being acquired for a collective (e.g., $150,000). Designers will need to weigh which guidelines are most critical, or how they need to be modified, in order to balance the human-collective system complexity and operation.

The design guidance inspired from the single operator-collective evaluation analyses provided recommendations that can be implemented into human-collective systems that have been assessed in great depth with respect to visualizations, models, and control mechanisms. The biologically inspired design guidance expands on the single operator-collective inspired guidance by identifying other characteristics that can be applied in order to address additional research questions for transparency in human-collective systems. Both sets of guidelines must be investigated further in order to understand how real world use scenario limitations influence the reliability of the recommendations. The design guidelines suggest how to mitigate challenging issues within robotic collective systems and offers the opportunity to explore these behaviors further in future work. Providing design guidance for human-collective systems will begin to create standards within the field.

Chapter 6: Conclusion

Collective robotic systems are biologically-inspired and composed of many simple individual entities that exhibit behaviors found in spatial swarms (e,g,, fish), colonies (e.g., ants), or a combination of both (e.g., bees). Collective robotic systems are advantageous due to their apparent global intelligence and emergent behaviors. Many applications can benefit from the incorporation of collectives, including environmental monitoring, disaster response missions, and infrastructure support. Designers of human-collective systems continue to debate what system design elements (e.g., models, visualizations, and control mechanisms) are needed in order to provide transparency of collective behaviors and enable operators to positively influence collectives. Integrating transparency into the system can mitigate poor operator behaviors, help attain meaningful and insightful information exchanges between the operator and collective, enable positive operator influence on collectives, and improve the human-collective's overall effectiveness and performance. Few human-collective evaluations have been conducted, many of which have only assessed how one system design element may impact human-collective behaviors, such as the human-collective performance.

This dissertation developed a transparency definition for collective systems [2] that was leveraged to assess how to achieve transparency in a single human-collective system. Two models, one consensus decision-making model and another that required operator influence in order to achieve the task, and two visualizations, a traditional and abstract collective representation, were evaluated for a sequential best-of-n decision-making task with four collectives, each consisting of 200 individual collective entities.

Transparency was evaluated with respect to how the model and visualization impacted 1) human operators who possess different capabilities, 2) operator comprehension, 3) system usability, and 4) human-collective performance. The specific transparency factors associated with each of these four categories were identified and evaluated in order to provide insight about the transparency implications. Transparency design guidance was created in order to aid the design of future human-collective systems. One set of guidelines were inspired from the results and discussions of the single human-collective analyses and another set was based on a review of the biological literature.

The models and visualizations provided transparency differently independently. There were advantages and disadvantages associated with both model and visualization types. The first visualization analysis determined that the abstract visualization provided the best transparency, since operators with different individual capabilities were able to perform relatively the same and the human-collective team performed better. The second analysis considered the affects of both the models with the visualizations and determined that no single model and visualization combination provided the best transparency, rather advantages and disadvantages associated with both models, visualizations, and particular model with visualization combinations were identified. The different outcomes between the two analyses suggest that transparency cannot be quantified by using the best system design elements, but instead must be quantified by considering how the transparency of the different system design elements, including the models (i.e., algorithms), interact with one another and how that system transparency influences human-collective interactions and performance.

This dissertation has made novel contributions to the collective robotics domain by defining transparency for human-collective systems, providing insight about how to achieve transparency in these systems, evaluating transparency embedded in multiple

system design elements of a single human-collective system, as well as providing design guidance for future systems. Specific contributions are described in Chapter 6.1, and the opportunities for future work are discussed in Chapter 6.2.

6.1 Contributions

There were four primary contributions generated by this dissertation.

1. *Transparency definition for human-collective systems.* This dissertation created the first transparency definition for human-collective systems that leveraged a commonly used robotics transparency definition and factors from traditional human-machine and human-robot domains [2]. The transparency definition identifies pertinent information to provide to the operator and collectives, and identifies methods to embed or promote transparency into the system. Two secondary contributions transpired from creating the human-collective system transparency definition. A human-machine system transparency factor concept map, directly applicable to human-collective systems, was created in order to aid designers in clarifying and identifying what factors are associated with transparency, what the relationships are between factors, and how those transparency factors are either influenced by transparency or affect transparency. Understanding which factors are associated with transparency are useful in determining which metrics are needed to assess transparency. Methods to embed and promote transparency in human-collective systems were identified in order to aid human-collective designers: 1) embed transparency via system features, which is the primary method explored in existing human-collective literature, 2) promote transparency by designing human-collective systems using specific guidelines, only the Gestalt prin-

ciples have been used in one evaluation [37], and 3) train the human operators and system, which was recently investigated in one evaluation [183].

2. *Metrics to assess transparency from a single human-collective evaluation.* Only two transparency human-collective system evaluations [29, 56] using colony based systems exist in the literature. This dissertation evaluated another colony based system and did a comparative analysis with that of Cody et al.'s [57], by assessing how different models (i.e., algorithms) and visualizations influenced human-collective interactions and behavior. A total of 37 metrics (excluding those in the Appendices) were evaluated, of which 27 were newly defined as a contribution of this dissertation. These metrics were used to determine which metrics effectively assessed transparency for human-collective systems. The existing transparency human-collective system literature in Chapters 2.0.1.3, 2.0.1.4, 2.0.2.3, 2.0.2.4, and 2.0.3.2, had only introduced approximately 15 metrics, most of which are performance related. Nine of the metrics defined in this dissertation were similar to those used in the existing human-collective literature, while five metrics were inspired from other human-machine domains and used to evaluate human-collective interactions and performance. Eleven new metrics were created and twelve were modified substantially from metrics in the human-machine interaction domains. Human-collective system evaluations that are performed in a simulation environment can use 23 of the metrics provided by this dissertation when performing an array of tasks. The number of useful metrics for in-situ evaluations is 14, although alterations of particular metrics can be used in real world scenarios, such as using eye-tracking data to assess the impacts of clutter and to permit assessing perception. This dissertation evaluated metrics that leveraged exist-

ing literature and human-machine domains in order to assess how transparency influenced operators with different individual capabilities (11 metrics), operator comprehension (13 metrics), system usability (24 metrics), and human-collective performance (16 metrics). Expanding the metrics to assess how transparency influenced behaviors other than human-collective performance will aid designers when assessing the impact of transparency related design decisions on human-collective systems.

3. *Quantification of human-collective system transparency.* Many of the existing transparency evaluations have only attempted to promote transparency and assess how human-collective behaviors were influenced by one system design element (e.g., visualization, model, or control mechanisms). This dissertation was the first evaluation to assess multiple system design elements in order to promote transparency, as well as determine how to quantify transparency for human-collective systems. Implementing the best system design elements together in one collective system design may not always promote the best transparency and yield optimal results. The human-collective system may become less transparent due to unanticipated and undesired operator behavior that results from the combined system design elements. Providing insight about how to quantify transparency in human-collective systems will provide a standard for determining the influence of design choices. Results from future human-collective systems will be more robust by assessing how multiple system design elements influence one another, as well as the human-collective interactions and behaviors.

4. *Transparency design guidance for human-collective systems.* Design guideline recommendations were created in order to inform designers how transparency can be

achieved for human-collective systems. Relationships between the design guidelines and the transparency factors were identified in order for designers to understand the implications of the guidance. No explicit design guidelines exist in the literature to aid designers of human-collective systems, although some work has identified concepts needed to design human-collective systems [184]. Providing guidance that can be used to inform design choices for models, visualizations, or control mechanisms in other systems will begin to create standards within the field. Many of the design guidelines leverage strategies used in other human-machine interaction domains, which may not always apply or scale to collectives that exhibit emergent behaviors in limited communication environments. Providing guidance to promote consistency among design principles will enable researchers to compare their findings with others, since current systems are very specific and cannot be generalized. The seven biologically-inspired behaviors that inspired creating 22 design guidelines can be applied in order to address a number of open research questions for transparency in human-collective systems. These biological behaviors have not been investigated extensively in this dissertation's evaluation or the existing literature. The biologically-inspired design guidelines suggest how to mitigate challenging issues within robotic collective systems and offers and opportunity to make these systems more robust.

6.2 Future Work

Interest in collective robotic systems will continue to increase due to the potential benefits that can be offered to operators, such as increased safety and support. Several potential research directions that investigate collective robotic systems and extend the

outcomes of this dissertation are provided.

Improve transparency factor concept map. This dissertation has shown the importance of the transparency factors for human-collective systems and how they either influence transparency or are affected by transparency. Further analysis can be conducted in order to improve the transparency factor concept map. More detailed definitions of the transparency factors and their corresponding relationships (i.e., links that connect the transparency factors together) are needed in order to reduce the ambiguity of what the factors mean and ensure that the relationships are accurate. Many of the relationship terminology is not insightful regarding the specific relationship between factors, such as "aspect of". There are opportunities to further investigate the relationships between factors. Some factors, such as observability, have multiple meanings in different domains, for example shared frame of reference for the automation and operator [39] versus easily perceived information that supports operators [185]. Some definitions will not be applicable to human-collective systems; therefore, as human-collective system research, grows the transparency concept map must be improved to reflect the findings from those evaluations and transition from a human-machine to a human-collective transparency concept map.

Transparency metrics reliability and repeatability. The metrics provided in this dissertation were used to successfully evaluate characteristics of human-collective system transparency; however, further analyses are needed in order to determine the reliability and repeatability of these metrics. Conducting a meta-analysis for metrics that were consistent across both the single human-collective evaluations can begin to provide insight about their reliability and reliability. More evaluations using these metrics will

help to inform their repeatability. Conducting these analyses will promote consistent metrics that can be used to assess transparency for future human-collective systems.

Improve and Expand the Bio-Inspired Design Guidelines. The following list of design guidance was inspired from the same biological behaviors discussed in Chapter 5.2. Further work remains to improve these particular design guidelines in order to provide meaningful information about how to embed these guidelines into human-collective systems, or incorporate them prior to system usage in order to promote transparency (e.g., training). These design guidelines, as well as those mentioned in Chapter 5.2, can be improved by investigating the recommended design strategies (e.g., representing state information, feedback, or suggestions) in more depth. Existing methodologies, such as using machine learning, or particular training protocols, can be leveraged to embedded the design guidance into the human-collective system. Comparisons can be made between the existing human-collective literature, not solely focused on transparency, and what is recommended from the design guidance in order to determine the validity of these recommendations.

1. Undesirable Emergent Behaviors.
    (a) Train operators how to use a respective system to mitigate undesirable emergent behaviors prior to system usage.
2. Individual Collective Entities Roles.
    (a) Train operators about the persistent *capabilities* and any individual collective entity roles before system usage.
3. Limited Communication Among Individual Collective Entities.

(a) Inform operators during training about limited communication between individual entities and how that limited communication may influence the development of collective behavior.

(b) Provide *feedback* about the communication latency to the operator.

4. Collective and Subgroup Information.

(a) Provide information to the operator about the collective or sub-group state.

5. Leadership.

(a) Provide control mechanisms to the operator that permit influencing the leaders.

(b) Implement model strategies to re-assign individuals to leadership roles if other leaders' capabilities fall below a critical threshold, such as low battery power, or other individual entities will provide more benefit to the collective serving as a leader, such as faster communication transmission.

6. Collective Influence of Individual Collective Entity Actions.

(a) Provide feedback to the operator about collective needs and if that need has impacted individual entities' actions.

(b) Provide suggestions about how to fulfill collective needs.

7. Feedback Loops.

(a) Provide control mechanisms to the operator that can be used multiple times, as long as previous actions are not negated, in order to increase support for particular behaviors.

(b) Implement a time limit that the control mechanism can influence individual collective entities in order to mitigate undesired behaviors.

(c) Implement control mechanisms that can attract or deter individual collective entities towards or away from desired locations.

Human-collective evaluations with imperfect communication. The single human-collective evaluation conducted as part of this dissertation assumed perfect communications between the human operator and the collective hub, as well as within the hub between the individual collective entities, like the majority of the evaluations from the existing literature. Real-world domain applications for collectives will not have perfect communication due to various factors, such as the communication modalities (e.g., WIFI or cellular coverage) limited bandwidth, or hardware and software limitations. Evaluations are needed to understand how human-collective systems transparency definition and factors, as well as the associated design guidance provided in this dissertation impact the human-collective teams' performance when systems have imperfect communication. Additional modifications may be needed, such as predictive collective state information, in order to achieve transparency.

Multiple human-collective evaluation. Collectives can be used as a shared resource to support multiple distributed human operators. The application of the transparency definition and factors, as well as the design guidance as applied to multiple human operators sharing collectives is important. Human organizational structures will influence human-collective teaming differently, since the organization is responsible for directing what, how, when, and why the human-collective team must perform particular tasks. Conflicts between the human organization, a single or multiple human operator(s), and

collective system may occur if expectations from the entities are not met. Identifying where deviations occur, whether it be by the organization, operator, or collective system, is necessary in order to design effective human-collective systems.

Bibliography

[1] Robert W. Proctor and Trisha Van Zandt. *Human Factors in Simple and Complex Systems*. CRC Press, Boca Raton, 2008.

[2] Karina A. Roundtree, Michael A. Goodrich, and Julie A. Adams. Transparency: Transitioning from human-machine systems to human-swarm systems. *Journal of Cognitive Engineering and Decision Making*, 13(3):171–195, 2019.

[3] Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. Swarm intell swarm robotics: a review from the swarm engineering perspectives. *Swarm Intelligence, Springer*, 7:1–41, 2013.

[4] Deborah M. Gordon. *Ants at work: how an insect society is organized*. Simon and Schuster, Oxford, UK, 1999.

[5] Thomas D. Seeley. *Honeybee democracy*. Princeton University Press, Princeton, NJ, 2010.

[6] Michele Ballerini, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, Irene Giardina, Vivien Lecomte, Alberto Orlandi, Greg Parisi, Andrea Procaccini, Massimiliano Viale, and Vladimir Zdrakovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1232–1237, 2008.

[7] Iain D. Couzin, Jens Krause, Nigel R. Franks, and Simon A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433:513–516, 2005.

[8] Gabriele Valentini, Eliseo Ferrante, and Marco Dorigo. The best-of-n problem in robot swarms: Formalization, state of the art, and novel perspectives. *Current Directions in Psychological Science*, 4:1–9, 2017.

[9] Levent Bayindir and Erol Sahin. A review of studies in swarm robotics. *Turkish Journal of Electrical Engineering & Computer Sciences*, 15:115–147, 2007.

[10] Andrew M. Hein, Sara Brin Rosenthal, George I. Hagstrom, Andrew Berdahl, Colin J. Torney, and Iain D. Couzin. The evolution of distributed sensing and collective computation in animal populations. *Ecology, Genomics and Evolutionary Biology*, 4:1–59, 2015.

[11] Mohammad Komareji and Roland Bouffanais. Resilience and controllability of dynamic collective. *PLoS One*, 46:1–15, 2013.

[12] Jean Scholtz. Theory and evaluation of human robot interactions. In *Hawaii International Conference on System Sciences*, page 10, 2003.

[13] Matthew D. Manning, Caroline E. Harriott, Sean T. Hayes, Julie A. Adams, and Adriane E. Seiffert. Heuristic evaluation of swarm metrics' effectiveness. In *ACM/IEEE Interational Conference on Human-Robot Interaction Extended Abstracts*, pages 17–18, 2015.

[14] David J. T. Sumpter. *Collective Animal Behavior*. Princeton University Press, Princeton, NJ, 2010.

[15] Ian Vine. Risk of visual detection and pursuit by a predator and the selective advantage of flocking behavior. *Journal of Theoretical Biology*, 30(2):405–422, 1971.

[16] Brian L. Patridge, Jonas Johansson, and John Kalish. The structure of schools of giant bluefin tuna in cape cod bay. *Environmental Biology of Fishes*, 9:253–262, 1983.

[17] Andrea Cavagna, Silvio M. Durate Queiros, Irene Giardina, Fabio Stefanini, and Massimiliano Viale. Diffusion of individual birds in starling flocks. *Proceedings of the Royal Society*, 280(1756):1–9, 2013.

[18] Ema Hensor, I. D. Couzin, R. James, and J. Krause. Modeling density-dependent fish shoal distributions in the laboratory and field. *Oikos*, 110(2):344–352, 2005.

[19] Matthew J. Hansen, Timothy M. Schaerf, and Ashley J. W. Ward. The influence of nutritional state on individual and group movement behavior in shoals of crimson-spotted rainbowfish (melanotaenia duboulayi). *Behavioral Ecology and Sociobiology*, 69(10):1713–1722, 2015.

[20] Kevin J. McGowan and Glen E. Woolfenden. A sentinel system in the florida scrub jay. *Animal Behavior*, 37(6):1000–1006, 1989.

[21] Roland W. Regeder and Jens Krause. Density dependence and numerosity in fright stimulated aggregation behavior of shoaling fish. *Philosophical Transactions of the Royal Society B*, 350(1334):381–390, 1995.

[22] Nicole Abaid and Maurizio Porfiri. Fish in a ring: Spatio-temporal pattern formation in one-dimensional animal groups. *Journal of the Royal Society*, 7:1441–1453, 2010.

[23] Stephen G. Reebs. Can a minority of informed leaders determine the foraging movements of a fish school? *Animal Behavior*, 59:403–409, 2000.

[24] George F. Young, Luca Scardovi, Andrea Cavagna, Irene Giardina, and Naomi E. Leonard. Starling flock networks manage uncertainty in consensus at low cost. *PLOS Computational Biology*, 9(1):1–7, 2013.

[25] Sean A. Rands, Guy Cowlishaw, Richard A. Pettifor, J. Marcus Rowcliffe, and Rufus A. Johnstone. Spontaneous emergence of leaders and follows in foraging pairs. *Nature*, 423:432–434, 2003.

[26] Jennifer S. Lewis, Douglas Wartzok, and Michael R. Heithaus. Highly dynamic fission-fusion species can exhibit leadership when traveling. *Behavioral Ecology and Sociobiology*, 65(5):1061–1069, 2011.

[27] Rolf G. Beilharz and Peter J. Mylrea. Social position an movement orders of dairy heifers. *Animal Behavior*, 11(4):529–533, 1963.

[28] Erol Sahin and William M. Spears. *Swarm Robotics*. Springer-Verlag, New York, NY, 2005.

[29] Jason R. Cody. *Discrete Consensus Decisions in Human-Collective Teams*. PhD thesis, Vanderbilt University, Vanderbilt University, Nashville, TN, USA, 2018.

[30] Julie A. Adams, Jessie Y. C. Chen, and Michael A. Goodrich. Swarm transparency. In *2018 ACM/IEEE International Conference on Human-Robot Interaction: Late-Breaking Reports*, pages 45–46, 2018.

[31] Iain D. Couzin, Jens Krause, Richard James, Graeme D. Ruxton, and Nigel R. Franks. Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology*, 218:1–11, 2002.

[32] Daniel T. Swain, Iain D. Couzin, and Naomi Ehrich Leonard. Coordinated speed oscillations in schooling killifish enrich social communication. *Journal of Nonlinear Science*, 25(5):1077–1109, 2015.

[33] John E. Treherne and William A. Foster. Group transmission of predator avoidance behavior in a marine insect: The trafalgar effect. *Animal Behavior*, 29:911–917, 1981.

[34] Phillip Walker, Steven Nunanally, Michael Lewis, Andreas Kolling, Nilanjan Chakraborty, and Katia Sycara. Neglect benevolence in human control of swarms in the presence of latency. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 3009–3014, 2012.

[35] Daniel S. Brown, Michael Goodrich, Shin-Young Jung, and Sean Kerman. Two invariants of human-swarm interaction. *Journal of Human-Robot Interaction*, 5:1–31, 2016.

[36] Ellen Haas, MaryAnne Fields, Susan Hill, and Christopher Stachowiak. Extreme scalability: Designing interfaces and algorithms for soldier-robotic swarm interaction. Technical Report Technical Report ARL-TR-4800, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, 2009.

[37] Sasanka Nagavalli, Shih-Yi Chen, Michael Lewis, Nilanjan Chakraborty, and Katia Sycara. Bounds of neglect benevolence in input timing for human interaction with robotic swarms. In *ACM/IEEE International Conference on Human-Robot Interactions*, pages 197–204, 2015.

[38] John Harvey, Kathryn Elizabeth Merrick, and Hussein A. Abbass. Assessing human judgment of computationally generated swarming behavior. *Frontiers in Robotics and AI*, 5, 2018.

[39] Jessie Y. C. Chen, Katelyn Procci, Michael Boyce, Julia L. Wright, Andre Garcia, and Michael J. Barnes. Situation awareness-based agent transparency. Technical Report Technical Report ARL-TR-6905, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, 2014.

[40] Phillip Walker, Christopher Miller, Joseph Mueller, Katia Sycara, and Michael Lewis. *A Playbook-Based Interface for Human Control of Swarms*, pages 61–88. CRC Press, Boca Raton, FL, 2019.

[41] Philip Walker, Michael Lewis, and Katia Sycara. The effect of display type on operator prediction of future swarm states. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 2521–2526, 2016.

[42] Karina A. Roundtree, Matthew D. Manning, and Julie A. Adams. Analysis of human-swarm visualizations. In *Human Factors and Ergonomics Society Annual Meeting*, page 287–291, 2018.

[43] Adriane E. Seiffert, Sean T. Hayes, Caroline E. Harriott, and Julie A. Adams. Motion perception of biological swarms. In *Annual Cognitive Science Society Meeting*, pages 2128–2133, 2015.

[44] Andreas Kolling, Steven Nunnally, and Michael Lewis. Towards human control of robot swarms. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 89–96, 2012.

[45] Shin-Young Jung and Michael A Goodrich. Multi-robot perimeter-shaping through mediator-based swarm control. In *International Conference on Advanced Robotics*, pages 1–6, 2013.

[46] Andreas Kolling, Katia Sycara, Steven Nunnally, and Michael Lewis. Human swarm interaction: An experimental study of two types of interaction with foraging swarms. *Journal of Human-Robot Interaction*, 2:103–128, 2013.

[47] Philip Walker, Steven Nunnally, Michael Lewis, Nilanjan Chakraborty, and Katia Sycara. Levels of automation for human influence of robot swarms. In *Human Factors and Ergonomics Society Annual Meeting*, pages 429–433, 2013.

[48] Deborah M. Gordon. *Ant encounters interaction networks and colony behavior*. Princeton University Press, Princeton, NJ, 2010.

[49] Nigel R. Franks and Tom Richardson. Teaching in tandem-running ants. *Nature*, 153:153–153, 2006.

[50] Stephan C. Pratt, Eamonn B. Mallon, David J. T. Sumpter, and Nigel R. Franks. Quorum sensing, recruitment, and collective decision-making during colony emigration by the ant lepthothrax albipennis. *Behavioral Ecology and Sociobiology*, 52(2):117–127, 2002.

[51] Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence*. Oxford University Press, New York, NY, 1999.

[52] Edward O. Wilson. The relation between caste ratios and division of labor in the ant genus *Pheidole Hymenoptera: Formicidae*. *Behavioral Ecology and Sociobiology*, 16:89–98, 1984.

[53] Bert Holldobler and Edward O. Wilson. *The Ants*. Cambridge University Press, Cambridge, MA, 1990.

[54] David J. T. Sumpter. The principles of collective animal behavior. *Philosophical Transactions of the Royal Society B*, 361:5–22, 2006.

[55] Jacob W. Crandall, Nathan Anderson, Chace Ashcraft, John Grosh, Jonah Henderson, Joshua McClellan, Aadesh Neupane, and Michael A. Goodrich. Human-swarm interaction as shared control: Achieving flexible fault-tolerant systems. *Engineering Psychology and Cognitive Ergonomics: Performance, Emotion and Situation Awareness*, 10275, 2017.

[56] C. Chace Ashcraft, Michael A. Goodrich, and Jacob W. Crandall. Moderating operator influence in human-swarm systems. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 4275–4282, 2019.

[57] Jason R. Cody, Karina A. Roundtree, and Julie A. Adams. Human-collective collaboration site selection. *arXiv:2004.09581*, 2020.

[58] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 2004.

[59] John D. Lee. Trust, trustworthiness, and trustability. In *Presentation at the Workshop on Human Machine Trust for Robust Autonomous Systems*, volume 31, 2012.

[60] Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[61] Mica R. Endsley, Betty Bolté, and Debra G. Jones. *Designing for Situation Awareness: An approach to user-centered design*. Taylor and Francis, London, 2003.

[62] Anand S. Rao and Michael P. Georgeff. Bdi agents: From theory to practice. In *International Conference on Multiagent Systems*, pages 312–319, 1995.

[63] Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael J. Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomic Science*, 19(3):259–282, 2018.

[64] Joseph E. Mercado, Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. Intelligent agent transparency in human-agent teaming for multi-uxv management. *Human Factors*, 58(3):401–415, 2016.

[65] Ryan W. Wohleber, Kimberly Stowers, Jessie Y. C. Chen, and Michael J. Barnes. Effects of agent transparency and communication framing on human-agent teaming. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 3427–3432, 2017.

[66] Dietmar Nedbal, Andreas Auinger, and Alexander Hochmeier. Addressing transparency, communication and participation in enterprise 2.0 projects. *Procedia Technology*, 9:676–686, 2013.

[67] Anthony R. Selkowitz, Shan G. Lakhmani, Cintya N. Larios, and Jessie Y. C. Chen. Agent transparency and the autonomous squad member. In *Human Factors and Ergonomics Society Annual Meeting*, pages 1319–1323, 2016.

[68] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. Shaping trust through transparent design: Theoretical and experimental guidelines. In P. Savage-Knepshield and J. Chen, editors, *Advances in Human Factors in Robots*

and Unmanned Systems. Advances in Intelligent Systems and Computing, volume 499, pages 127–126. Springer, Cham, Switzerland, 2017.

[69] Joseph B. Lyons. Being transparent about transparency: A model for human-robot interaction. In *Trust and Autonomous Systems: Association or the Advancement of Artificial Intelligence*, pages 48–53, 2013.

[70] Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3), 2017.

[71] Anthony R. Selkowitz, Shan G. Lakhmani, Jessie Y. C. Chen, and Michael Boyce. The effects of agent transparency on human interaction with an autonomous robotic agent. In *Human Factors and Ergonomics Society Annual Meeting*, pages 806–810, 2015.

[72] Scott Ososky, Tracey Sanders, Florian Jentsch, Peter Hancock, and Jessie Y. C. Chen. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *The International Society of for Optical Engineering*, volume 9084, pages 1–12, 2014.

[73] Taemie Kim and Pamela Hinds. Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *IEEE International Symposium on Robot and Human Interactive Communication*, pages 80–85, 2006.

[74] Sylvian Bruni, Jessica J. Marquez, Amy Brzezinski, Carl Nehme, and Yves Boussemart. Introducing a human-automation collaboration taxonomy (hact) in command and control decision-support systems. In *International Command and Control Research and Technology Symposium*, pages 1–13, 2007.

[75] Gloria Mark and Alfred Kobsa. The effects of collaborative and system transparency on cive usage: An empirical study and model. *Presence: Teleoperators & Virtual Environments*, 12:60–80, 2005.

[76] Tracy L. Sanders, Tarita Wixon, K. Elizabeth Schafer, Jessie Y. C. Chen, and P. A. Hancock. The influence of modality and transparency on trust in human-robot interaction. In *IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, pages 156–159, 2014.

[77] Elizabeth Kaltenbach and Igor Dolgov. On the dual nature of transparency and reliability: Rethinking factors that shape trust in automation. In *Human Factors and Ergonomics Society Annual Meeting*, pages 308–312, 2017.

[78] Kristina Höök. Steps to take before intelligent user interfaces become real. *Interacting with Computers*, 12:409–426, 2000.

[79] Jonathan Vitale, Meg Tonkin, Sarita Herse, Suman Ojha, Jesse Clark, Mary-Anne Williams, Xun Wang, and William Judge. Be more transparent and users will like you: A robot privacy and user experience design experiment. In *ACM/IEEE International Conference on Human-Robot Interaction*, page 379–387, 2018.

[80] Gene M. Alarcon, Rose Gamble, Sarah A. Jessup, Charles Walter, Tyler J. Ryan, David W. Wood, and Chris S. Calhoun. Application of the heuristic-systematic model to computer code trustworthiness: The influence of reputation and transparency. *Cogent Psychology*, 4:1–22, 2017.

[81] Tove Helldin. *Transparency for Future Semi-Automated Systems: Effects of Transparency on Operator Performance, Workload and Trust*. PhD thesis, Orebro University, Orebro University, 2014.

[82] Robert H. Wortham and Andreas Theodorou. Robot transparency, trust and utility. *Connection Science*, 29(3), 2017.

[83] Asaf Degani, Claudia V. Goldman, Omer Deutsch, and Omer Tsimhoni. On human-machine relations. *Cognition, Technology & Work*, 19:211–231, 2017.

[84] Amor Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. Measurement of trust in human-robot collaboration. In *International Symposium on Collaborative Technologies and Systems*, pages 106–114, 2007.

[85] Munjal Desai, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *ACM/IEEE international conference on Human-robot interaction*, pages 251–258, 2013.

[86] Jessie Y. C. Chen, Michael J. Barnes, Anthony R. Selkowitz, and Kimberly Stowers. Effects of agent transparency on human-autonomy teaming effectiveness. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 1838–1843, 2016.

[87] Anthony R. Selkowitz, Shan G. Lakhmani, and Jessie Y. C. Chen. Using agent transparency to support situation awareness of the autonomous squad member. *Cognitive Systems Research*, 46:13–25, 2017.

[88] Kimberly Stowers, Nicholas Kasdaglis, Olivia Newton, Shan G. Lakhmani, Ryan Wohleber, and Jessie Y. C. Chen. Intelligent agent transparency: The design and evaluation of an interface to facilitate human and intelligent agent collaboration. In *Human Factors and Ergonomics Society Annual Meeting*, pages 1706–1710, 2016.

[89] Jeff Rubin and Dana Chisnell. *Handbook of Usability Testing, Second Edition: How to Plan, Design and Conduct Effective Tests*. Wiley Publishing, Inc., 2008.

[90] Jakob Nielsen. *Usability Engineering*. Academic Press, Indianapolis, IN, 1993.

[91] David D. Woods and Erik Hollnagel. *Joint Cognitive Systems: Patterns in Cognitive Engineering*. Taylor & Francis Group, Boca Raton, FL, 2006.

[92] Robert Truxler, Emilie Roth, Ronald Scott, Stephen Smith, and Jeffery Wampler. Designing collaborative automated planners for agile adaptation to dynamic change. In *Human Factors and Ergonomics Society Annual Meeting*, pages 223–227, 2012.

[93] Thomas B. Sheridan and William L. Verplank. Human and computer control of undersea teleoperators. Technical report, Massachusetts Institute of Technology, Cambridge, MA, 1978.

[94] Jon P. Gant and Diana Burley Gant. Web portal functionality and state government e-service. In *Hawaii International Conference on System Sciences*, pages 1–10, 2002.

[95] Anas Ratib Al-Soud and Keiichi Nakata. Evaluating e-government websites in jordan: Accessibility, usability, transparency and responsiveness. In *IEEE International Conference on Progress in Informatics and Computing*, pages 761–765, 2010.

[96] Alexander J. DeWitt and Jasna Kuljis. Aligning usability and security: A usability study of polaris. In *Symposium on Usable privacy and security*, pages 1–7, 2006.

[97] Scott Ruoti, Brent Roberts, and Kent Seamons. Authentication melee: A usability analysis of seven web authentication systems. In *International Conference on World Wide Web*, pages 916–926, 2015.

[98] Cynthia Breazeal, Cory D. Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 383–388, 2005.

[99] Sigal Berman and Tzvi Ganel. Perception and action in remote and virtual environments. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 67–68, 2018.

[100] Jessie Y. C. Chen, Michael J. Barnes, Anthony R. Selkowitz, Kimberly Stowers, Shan G. Lakhmani, and Nicholas Kasdaglis. Human-autonomy teaming and agent transparency. In *International Conference on Intelligent User Interfaces*, pages 28–31, 2016.

[101] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. Introduction to special forum: Not so different after all: A cross-discipline view of trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 2004.

[102] John. D. Lee and N. Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.

[103] Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995.

[104] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *Australasian Conference on Information Systems*, pages 31–33, 2000.

[105] René F. Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Conference on Human Factors in Computing Systems*, pages 2390–2395, 2016.

[106] Joachim Meyer, Rebecca Wiczorek, and Torsten Günzler. Measures of reliance and compliance in aided visual scanning. *Human Factors*, 56(5):840–849, 2014.

[107] Stephen L. Jones and Priti Pradhan Shah. Diagnosing the locus of trust: A temporal perspective for trustor, trustee, and dyadic influences on perceived trustworthiness. *Journal for Applied Psychology*, 10(3):392–414, 2015.

[108] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. Effects of changing reliability on trust of robot systems. In *ACM/IEEE international conference on Human-Robot Interaction*, pages 73–80, 2012.

[109] Julian Sanchez. *Factors that Affect Trust and Reliance on an Automated Aid*. PhD thesis, Georgia Institute of Technology, Georgia Institute of Technology, 2006.

[110] Joseph B. Lyons, Matthew A. Clark, Alan R. Wagner, and Matthew J. Schuelke. Certifiable trust in autonomous systems: Making the intractable tangible. *AI Magazine*, 38(3):37–49, 2017.

[111] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

[112] Hasmik Atoyan, Jean-Rémi Duquet, and Jean-Marc Robert. Trust in new decision aid systems. In *Conference on l'Interaction Homme-Machine*, pages 115–122, 2006.

[113] Scott Ruoti, Nathan Kim, Ben Burgon, Timothy van der Horst, and Ken Seamons. Confused johnny: When automatic encryption leads to confusion and mistakes. In *Symposium on Usable Privacy and Security*, pages 1–19, 2013.

[114] Joseph B. Lyons, Kolina S. Koltai, Nhut T. Ho, Walter B. Johnson, David E. Smith, and R. Jay Shively. Engineering trust in complex automated systems. *Ergonomics in Design: The Quarterly of Human Factors Applications*, 24(1), 2016.

[115] Michael W. Boyce, Jessie Y. C. Chen, Anthony R. Selkowitz, and Shan G. Lakhmani. Effects of agent transparency on operator trust. In *ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 179–180, 2015.

[116] Jessie Y. C. Chen, Anthony R. Selkowitz, Kimberly Stowers, Shan G. Lakhmani, and Michael J. Barnes. Human-autonomy teaming and agent transparency. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 91–92, 2017.

[117] Aya Hussein, Sondoss Elsawah, and Hussein A. Abbass. The reliabiliy and transparency bases on trust in human-swarm interaction: Principles and implications. *Ergonomics*, pages 1–17, 2020.

[118] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Planning with trust for human-robot collaboration. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 307–315, 2018.

[119] Shih-Yi Chien, Michael Lewis, Katia Sycara, Asiye Kumuru, and Jyi-Shane Liu. Influence of culture, transparency, trust, and degree of automation on automation use. *IEEE Transactions on Human-Machine Systems*, 50(3):205–214, 2020.

[120] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable planning. *arXiv:1709.10256v1*, 2017.

[121] Bradley Hayes and Julie A. Shah. Improving robot controller transparency through autonomous policy explanation. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 303–312, 2017.

[122] Frode Sormo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning - perspectives and goals. *Artificial Intelligence Review*, 24:109–143, 2005.

[123] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[124] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. The effect of explanations on perceived control and behaviors in intelligent systems. In *Extended Abstracts on Human Factors in Computing Systems*, pages 181–186, 2013.

[125] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Conference on Human Factors in Computing Systems*, pages 2119–2128, 2009.

[126] Michael W. Floyd and David W. Aha. Using explanations to provide transparency during trust-guided behavior adaptation. *AI Communications*, 30:281–294, 2017.

[127] Daniel Carrillo-Zapata, Emma Milner, Julian Hird, Georgios Tzoumas, Paul J. Vardanega, Mahesh Sooriyabandara, Manuel Giuliani, Alan F. T. Winfield, and Sabine Hauert. Mutual shaping in swarm robotics: User studies in fire and rescue, storage organization, and bridge inspection. *Frontiers in Robotics and AI*, 7(53):1–19, 2020.

[128] Alexandra Kirsch, Thibault Kruse, E. Akin Sisbot, Rachid Alami, Martin Lawitzky, Drazen Brscic, Sandra Hirche, Patrizia Basili, and Stefan Glasauer. Plan-based control of joint human-robot activities. *Kunstliche Intelligenz*, 24(3), 2010.

[129] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. Timing is key for robot trust repair. In *International Conference on Social Robotics*, pages 574–583, 2015.

[130] Ning Wang, David V. Pynadath, and Susan G. Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *ACM/IEEE International Conference on Human Robot Interaction*, pages 109–116, 2016.

[131] Ning Wang, David V. Pynadath, and Susan G. Hill. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 997–1005, 2016.

[132] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000.

[133] Haydee M. Cuevas, Stephen M. Fiore, Barrett S. Caldwell, and Laura Strater. Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviation, Space, and Environmental Medicine*, 78(5):63–70, 2007.

[134] Mark St. John, Harvey S. Smallman, and Daniel I. Manes. Recovery from interruptions to a dynamic monitoring task: The beguiling utility of instant replay. In *Human Factors and Ergonomics Society Annual Meeting*, pages 473–477, 2005.

[135] Harvey S. Smallman and Mark St. John. Chex (change history explicit): New hci concepts for change awareness. In *Human Factors and Ergonomics Society Annual Meeting*, pages 528–532, 2003.

[136] Phillip Walker, Steven Nunnally, Michael Lewis, Andreas Kolling, Nilanjan Chakraborty, and Katia Sycara. Neglect benevolence in human control of swarms in the presence of latency. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 3009–3014, 2012.

[137] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58:697–718, 2003.

[138] Ericka Rovira, Austin Cross, Evan Leitch, and Craig Bonacento. Displaying contextual information reduces the costs of imperfect decision automation in rapid retasking of isr assets. *Human Factors*, 56(6):1036–1049, 2014.

[139] Eileen B. Entin, Elliott E. Entin, and Daniel Serfaty. Optimizing aided target-recognition performance. In *Human Factors and Ergonomics Society Annual Meeting*, pages 233–237, 1996.

[140] John E. Mathieu, Tonia S. Heffner, Gerald F. Goodwin, Eduardo Salas, and Janis A. Cannon-Bowers. The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85(2):273–283, 2000.

[141] Jens Rasmussen. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(3):257–266, 1983.

[142] L. Richard Ye and Paul E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, 19(2):157–172, 1995.

[143] O. Can Gorur, Benjamin Rosman, Fikret Sivrikaya, and Sahin Albayrak. Social cobots: Anticipatory decision-making for collaborative robots incorporating unexpected human behaviors. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 398–406, 2018.

[144] Karl E. Weick. *Sensemaking in Organizations*. Thousand Oaks : Sage Publications, Thousand Oaks, CA, 1995.

[145] Cheryl A. Bolstad, Haydee M. Cuevas, Cleotilde Gonzalez, and Mike Schneider. Modeling shared situation awareness. In *Conference on Behavior Representation in Modeling and Simulation*, pages 1–8, 2005.

[146] Stephen M. Fiore, Florian Jentsch, Irma Becerra-Fernandez, Eduardo Salas, and Neal Finkelstein. Integrating field data with laboratory training research to improve the understanding of expert human-agent teamwork. In *Hawaii International Conference on System Sciences*, pages 1–11, 2005.

[147] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *AAAI Conference on Artificial Intelligence*, pages 4762–4763, 2017.

[148] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, Birna van Riemsdijk, and Maarten Sierhuis. The fundamental principle of coactive design: Interdependence must shape autonomy. In *Coordination, Organization, Institutions, and Norms in Agent Systems VI*, pages 172–191, 2011.

[149] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 319–326, 2016.

[150] Christopher D. Wickens, John D. Lee, Yili Liu, and Sallie E. Gordon Becker. *An Introduction to Human Factors Engineering*. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

[151] Ilona Roth and John P. Frisby. *Perception and representation: a cognitive approach*. Open University Press, Philadelphia, PA, 1986.

[152] Thomas F. Shipley and Philip J. Kellman. *From fragments to objects: Segmentation and Grouping in Vision*. Elsevier Science, Amsterdam, The Netherlands, 2001.

[153] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, 2001.

[154] Catherine M. Burns and John R. Hajdukiewicz. *Ecological Interface Design*. CRC Press, Boca Raton, FL, 2004.

[155] Neville Moray. Mental models in theory and practice. *Attention and Performance*, 17:222–258, 1999.

[156] John M. Carroll and Judithm Reitman Olson. *Mental models in human-computer interaction: Research issues about what the user of software knows*. National Academy Press, Washington D.C., USA, 1987.

[157] Christopher D. Wickens, Justin G. Hollands, Simon Banbury, and Raja Parasuraman. *Engineering Psychology and Human Performance*. Pearson, Boston, fourth edition, 2013.

[158] Marie-Eve Jobidon, Alexandra Muller-Gass, Matthew Duncan, and Ann-Renee Blais. The enhancement of mental models and its impact on teamwork. In *Human Factors and Ergonomics Society Annual Meeting*, pages 1703–1707, 2012.

[159] James Reason. *Human Error*. Cambridge University Press, New York, NY, 1990.

[160] Stanley N. Roscoe. Airborne displays for flight and navigation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 10(4):321–332, 1968.

[161] Paul R. Rosenbaum. *Design of Observational Studies*. Springer, New York, NY, 2010.

[162] Donald A. Norman. *Some observations on mental models*. Morgan Kaufmann Publishers Inc., New York, NY, 1983.

[163] Jenny Preece and Yvonne Rogers. *Interaction Design: Beyond Human-computer Interaction*. Wiley, West Sussex, England, 2007.

[164] Robert M. Yerkes and John D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology of Psychology*, 18:459–482, 1908.

[165] Martin Voshell, David D. Woods, and Flip Phillips. Overcoming the keyhole in human-robot coordination: Simulation and evaluation. In *Human Factors and Ergonomics Society Annual Meeting*, pages 442–446, 2005.

[166] Alan D. Baddeley. *Working memory*. Oxford University Press, 1986.

[167] Christopher D. Wickens and Jason S. McCarley. *Applied attention theory*. CRC Press, Boca Raton, FL, 2008.

[168] Dennis Andersson, Amy Rankin, and Darryl Diptee. Approaches to team performance assessment: A comparison of self-assessment reports and behavioral observer scales. *Cognition, Technology & Work*, 19, 2017.

[169] Robert A. Bjork. Assessing our own competence: Heuristics and illusions. *Attention and Performance*, pages 435–459, 1999.

[170] Andreagiovanni. Reina, Gabriele Valentini, Cristian Fernandez-Oto, Marco Dorigo, and Vito Trianni. A design pattern for decentralised decision making. *PLoS ONE*, 10(10):1–18, 2015.

[171] Karina A. Roundtree, Jason R. Cody, Jennifer Leaf, H. Onan Demirel, and Julie A. Adams. Visualization design for human-collective teams. In *Human Factors and Ergonomics Society Annual Meeting*, pages 417–421, 2019.

[172] Steven G. Vandenberg and Allan R. Kuse. Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47(2):599–604, 1978.

[173] Randall W. Engle. Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11(1):19–23, 2002.

[174] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, pages 139–183, 1988.

[175] Steven J. Selcon, R. M. Taylor, and Eva Koritsas. Workload or situational awareness? tlx vs. sart for aerospace systems design evaluation. In *Human Factors and Ergonomics Society Annual Meeting*, pages 62–66, 1991.

[176] Douglas J. Gillan, Kritina Holden, Susan Adam, Marianne Rudisill, and Laura Magee. How should fitts' law be appled to human-computer interaction? *Interacting with Computers*, 4(3):291–313, 1992.

[177] Karina A. Roundtree, Jason R. Cody, Jennifer Leaf, H. Onan Demirel, and Julie A. Adams. Human-collective visualization transparency. *arXiv:2003.10681*, 2020.

[178] John P. Chin, Virginia A. Diehl, and Kent L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Conference on Human Factors in Computing Systems*, pages 213–218, 1988.

[179] Russell Golman, David Hagmann, and John H. Miller. Polya's bees: A model of decentralized decision-making. *Science Advances*, 1(8):1–7, 2015.

[180] George A. Miller. The magical number seven, plus or minus two: Some limits on capacity for processing information. *Psychological Review*, 101(2):343–352, 1994.

[181] Musad Haque, Christopher Ren, Electa A. Baker, Douglas Kirkpatrick, and Julie A. Adams. Analysis of swarm communication models. In *International Workshop on Combinations of Intelligent Methods and Applications*, 2016.

[182] Rita Borgo, Johannes Kehrer, David H. S. Chung, Eamonn Maguire, Robert S. Laramee, Helwig Hauser, Matthew Ward, and Min Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics 2013 - State of the Art Reports*, pages 39–63. The Eurographics Association, 2013.

[183] Jonas D. Hasbach, Thomas E. F. Witte, and Maren Bennewitz. On the importance of adaptive operator training in human-swarm interaction. In *2012 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3009–3014, 2012.

[184] Andreas Kolling, Phillip Walker, Nilanjan Chakraborty, Katia Sycara, and Michael Lewis. Human interaction with robot swarms: A survey. *IEEE Transactions on Human-Machine Systems*, 46:9–26, 2015.

[185] Nadine B. Sarter and David D. Woods. Team play with a powerful and independent agent: Operational experiences and automation surprises on the airbus a-320. *The Journal of the Human Factors and Ergonomics Society*, 39(4):553–569, 1997.

APPENDICES

Appendix A: Demographic Questionnaire

The demographic questions used in both evaluations are identified in Table A.1. The participants were asked to circle one answer for each question.

Table A.1: Demographic questionnaire.

Age:	18-30	31-40	41-50	51-60	60+		
Gender:	Male			Female			
Education:	High School	Some College	College Degree	Some Graduate School	Masters	Doctorate	
How many hours per week do you use a desktop, or laptop computer with either a computer mouse or laptop trackpad?	0 hrs	$<$ 3hrs	3-8 hrs	$>$8 hrs			
Please indicate your proficiency with video games from 1 to 7, where 1 is little to no proficiency and 7 is highly proficient.	1	2	3	4	5	6	7

Appendix B: SA Probe Questions

The SA probe questions used in the IA evaluation are identified in Table B.1. The questions were tailored to the current situation based on the stochasticity of the evaluation.

Table B.1: SA probe questions.

Trial	Time	Question	Lvl	
1a	0:50	What collectives are investigating Target __?	1	1 2 3 4
	1:50	What target is collective __ likely to choose?	3	
	2:50	Which collective has achieved a majority of for Target __?	2	1 2 3 4
	3:50	Is Target __ in being investigated by any collectives?	1	Yes No
	4:50	What is the highest valued target available to collective __?	1	
	5:50	Which behavior are most of the agents in collective __ doing?	2	U F C X
1b	0:50	Which collective has the highest support for Target __?	1	1 2 3 4
	1:50	Which is the highest valued target for collective __?	1	
	2:50	Which collective is primarily exploring?	2	1 2 3 4
	3:50	Is Target __ in range of collective __?	2	Yes No
	4:50	Which collective will make the next decision?	3	1 2 3 4
	5:50	Will support for Target __ decrease?	3	Yes No
2a	0:50	Is Target __ likely to be selected by collective __?	3	Yes No
	1:50	Which target has the highest value for collective __?	1	
	2:50	In collective __, which of the four behaviors are most of the agents performing?	2	U F C X
	3:50	Is target __ in range of both collective __ and collective __?	2	Yes No
	4:50	Which collective is closest to achieving a majority for a target?	3	1 2 3 4
	5:50	Which Collective or collectives are investigating Target __?	1	1 2 3 4

Trial	Time	Question	Lvl	
2b	0:50	Is target __ being investigated by multiple collectives?	1	Yes No
	1:50	Which target should not be chosen by collective __?	2	
	2:50	Is collective __ investigating Target __?	1	Yes No
	3:50	Which collective has the highest support for Target __?	1	
	4:50	Is an agent in collective __ more likely to be exploring than an agent in collective __?	2	Yes No
	5:50	Which collective is closest to picking an optimal site?	3	1 2 3 4

Appendix C: Post-Trial Questionnaire

The post-trial questions used in both evaluations are identified in Table C.1. The participants were asked to respond to the following prompts regarding their experience during the experiment.

Table C.1: Post-trial questionnaire.

Please indicate the effectiveness of each command for controlling the collective decision in support of the highest value target from 1 to 7, where 1 is not effective and 7 is very effective.	Investigate	1	2	3	4	5	6	7
	Abandon	1	2	3	4	5	6	7
	Decide	1	2	3	4	5	6	7
I was able to understand the collective's behavior during each decision this trial from 1 to 7, where 1 is never and 7 is always.		1	2	3	4	5	6	7
The collective chose the target I felt was the best in each decision during this trial from 1 to 7, where 1 is never and 7 is always.		1	2	3	4	5	6	7

Appendix D: Post Experiment Questionnaire

The post-experiment questions used in the IA evaluation are identified in Table D.1. The participants were asked to rank the collective behavior algorithms, best or worst, in each of the questions below.

Table D.1: Post-experiment questionnaire.

		Best	Worst
Please rank collective responsiveness to your requests according to each collective trial.	Trial 1		
	Trial 2		
Please rank your ability to choose the highest quality target with each collective trial.	Trial 1		
	Trial 2		
Please rank your understanding of the collective's behavior in each trial.	Trial 1		
	Trial 2		

Appendix E: Additional Operator Comprehension Data

The sum of collective and target left- and right-clicks an operator made was the total number of *interactions*. The number of interactions 15 seconds before asking, while

Table E.1: Interactions descriptive statistics 15 seconds before asking, while asking, and during response to SA probe question by SA level using the M_2 model.

	Timing	SA Level	Mean (SD)	Median (Min/Max)
IA	Before	SA_O	5.52 (3.91)	5 (0/22)
		SA_1	5.69 (3.91)	5 (0/22)
		SA_2	5.12 (3.81)	4 (0/18)
		SA_3	5.77 (4.05)	5 (0/16)
	Asking	SA_O	1.18 (1.22)	1 (0/5)
		SA_1	1.14 (1.15)	1 (0/4)
		SA_2	1.21 (1.25)	1 (0/5)
		SA_3	1.22 (1.3)	1 (0/5)
	Responding	SA_O	4.37 (3.24)	4 (0/15)
		SA_1	3.92 (2.93)	3.5 (0/13)
		SA_2	4.38 (3.21)	4 (0/14)
		SA_3	5.15 (3.67)	4.5 (0/15)
Collective	Before	SA_O	5.75 (3.5)	5.5 (0/18)
		SA_1	5.77 (3.6)	5 (0/18)
		SA_2	5.71 (3.51)	6 (0/17)
		SA_3	5.77 (3.31)	6 (0/14)
	Asking	SA_O	1.97 (1.62)	2 (0/9)
		SA_1	1.68 (1.45)	1 (0/7)
		SA_2	2.28 (1.72)	2 (0/9)
		SA_3	2.11 (1.72)	2 (0/6)
	Responding	SA_O	4.32 (2.64)	4 (0/13)
		SA_1	3.96 (2.63)	4 (0/11)
		SA_2	4.29 (2.57)	4 (0/11)
		SA_3	5.14 (2.62)	6 (0/13)

asking, and during response to a SA probe question descriptive statistics using the M_2 and M_3 models are presented in Tables E.1 and E.2, respectively. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables E.3 and E.4, respectively. The Spearman correlation between the interactions and SA probe accuracy are provided in Table E.5.

Table E.2: Interactions descriptive statistics 15 seconds before asking, while asking, and during response to SA probe question by SA level using the M_3 model.

	Timing	**SA Level**	**Mean (SD)**	**Median (Min/Max)**
IA	Before	SA_O	8.08 (4.7)	8 (0/26)
		SA_1	8.35 (4.12)	9 (0/20)
		SA_2	6.94 (4.76)	7 (0/19)
		SA_3	9.17 (5.25)	9 (0/26)
	Asking	SA_O	2.19 (1.72)	2 (0/9)
		SA_1	1.85 (1.29)	2 (0/5)
		SA_2	2.77 (2.08)	3 (0/9)
		SA_3	1.98 (1.66)	2 (0/7)
	Responding	SA_O	7.24 (3.71)	7 (0/18)
		SA_1	6.88 (3.47)	7 (0/18)
		SA_2	7.31 (3.49)	7 (0/15)
		SA_3	7.74 (4.33)	8 (0/18)
Collective	Before	SA_O	7.9 (4.69)	7.5 (0/28)
		SA_1	7.3 (4.84)	7 (0/17)
		SA_2	8.08 (4.43)	7 (0/28)
		SA_3	8.41 (4.88)	9 (0/18)
	Asking	SA_O	2.9 (2.03)	3 (0/10)
		SA_1	2.79 (1.98)	3 (0/8)
		SA_2	3.17 (2.19)	3 (0/10)
		SA_3	2.6 (1.77)	3 (0/8)
	Responding	SA_O	5.88 (3.29)	6 (0/18)
		SA_1	5.54 (3.18)	5 (0/14)
		SA_2	6.14 (3.09)	6 (0/15)
		SA_3	5.89 (3.71)	6 (0/18)

Table E.3: Within model comparison statistics (DOF = 1) of interactions 15 seconds before asking, while asking, and during response to SA probe question by SA level.

	Timing	SA Level	Sample Size	Mann-Whitney-Wilcoxin
IA	Before	SA_O	670	U = 37396, $\rho < 0.001$
		SA_1	281	U = 6067.5, $\rho < 0.001$
		SA_2	224	U = 4781, $\rho < 0.01$
		SA_3	165	U = 2072, $\rho < 0.001$
	Asking	SA_O	670	U = 36766, $\rho < 0.001$
		SA_1	281	U = 6714.5, $\rho < 0.001$
		SA_2	224	U = 3491, $\rho < 0.001$
		SA_3	165	U = 2510, $\rho < 0.01$
	Responding	SA_O	670	U = 30856, $\rho < 0.001$
		SA_1	281	U = 4931.5, $\rho < 0.001$
		SA_2	224	U = 3300.5, $\rho < 0.001$
		SA_3	165	U = 2200.5, $\rho < 0.001$
Collective	Before	SA_O	672	U = 40976, $\rho < 0.001$
		SA_1	266	U = 7093.5, $\rho = 0.01$
		SA_2	252	U = 5339, $\rho < 0.001$
		SA_3	154	U = 1985, $\rho < 0.001$
	Asking	SA_O	672	U = 41234, $\rho < 0.001$
		SA_1	266	U = 5816.5, $\rho < 0.001$
		SA_2	252	U = 5921, $\rho < 0.001$
		SA_3	154	U = 2453, $\rho = 0.07$
	Responding	SA_O	672	U = 40164, $\rho < 0.001$
		SA_1	266	U = 6066, $\rho < 0.001$
		SA_2	252	U = 4995, $\rho < 0.001$
		SA_3	154	U = 2551, $\rho = 0.16$

Table E.4: Between visualization comparison statistics (DOF = 1) of interactions 15 seconds before asking, while asking, and during response to SA probe question by SA level.

	Timing	SA Level	Sample Size	Mann-Whitney-Wilcoxin
M_2	Before	SA_O	670	U = 59648, ρ = 0.16
		SA_1	294	U = 11128, ρ = 0.63
		SA_2	224	U = 7090, ρ = 0.09
		SA_3	152	U = 2940, ρ = 0.8
	Asking	SA_O	670	U = 72110, $\rho < 0.001$
		SA_1	294	U = 13061, $\rho < 0.01$
		SA_2	224	U = 8634.5, $\rho < 0.001$
		SA_3	152	U = 3744.5, $\rho < 0.001$
	Responding	SA_O	670	U = 57961, ρ = 0.46
		SA_1	294	U = 11178, ρ = 0.58
		SA_2	224	U = 6390.5, ρ = 0.81
		SA_3	152	U = 3066.5, ρ = 0.47
M_3	Before	SA_O	672	U = 55082, ρ = 0.59
		SA_1	253	U = 6840, ρ = 0.07
		SA_2	252	U = 9088, ρ = 0.03
		SA_3	167	U = 3235, ρ = 0.42
	Asking	SA_O	672	U = 67980, $\rho < 0.001$
		SA_1	253	U = 10068, $\rho < 0.001$
		SA_2	252	U = 8665, ρ = 0.15
		SA_3	167	U = 4195.5, ρ = 0.02
	Responding	SA_O	672	U = 44771, $\rho < 0.001$
		SA_1	253	U = 6211, $\rho < 0.01$
		SA_2	252	U = 6261, $\rho < 0.01$
		SA_3	167	U = 2637, $\rho < 0.01$

Table E.5: Spearman correlation analysis between interactions and SA probe accuracy 15 seconds before asking, while asking, and during response to SA probe question by SA level.

	Timing	SA Level	IA Correlation	Collective Correlation
M_2	Before	SA_O	r = 0.08, ρ = 0.14	r = 0.07, ρ = 0.23
		SA_1	r = -0.002, ρ = 0.98	r = 0.02, ρ = 0.83
		SA_2	r = 0.21, ρ = 0.03	r = 0.03, ρ = 0.79
		SA_3	r = 0.1, ρ = 0.36	r = 0.24, ρ = 0.04
	Asking	SA_O	r = 0.04, ρ = 0.46	r = 0.09, ρ = 0.09
		SA_1	r = -0.09, ρ = 0.27	r = 0.11, ρ = 0.16
		SA_2	r = 0.14, ρ = 0.14	r = 0.1, ρ = 0.3
		SA_3	r = 0.14, ρ = 0.22	r = 0.09, ρ = 0.48
	Responding	SA_O	r = 0.17, $\rho < 0.01$	r = 0.05, ρ = 0.36
		SA_1	r = 0.14, ρ = 0.11	r = 0.008, ρ = 0.92
		SA_2	r = 0.2, ρ = 0.04	r = 0.01, ρ = 0.9
		SA_3	r = 0.17, ρ = 0.13	r = 0.2, ρ = 0.08
M_3	Before	SA_O	r = 0.08, ρ = 0.16	r = -0.01, ρ = 0.85
		SA_1	r = 0.14, ρ = 0.1	r = -0.11, ρ = 0.25
		SA_2	r = 0.13, ρ = 0.18	r = 0.04, ρ = 0.67
		SA_3	r = -0.03, ρ = 0.79	r = 0.09, ρ = 0.44
	Asking	SA_O	r = 0.05, ρ = 0.35	r = 0.1, ρ = 0.07
		SA_1	r = 0.14, ρ = 0.1	r = -0.03, ρ = 0.73
		SA_2	r = 0.18, ρ = 0.06	r = 0.06, ρ = 0.47
		SA_3	r = -0.19, ρ = 0.09	r = 0.2, ρ = 0.07
	Responding	SA_O	r = 0.08, ρ = 0.12	r = -0.02, ρ = 0.75
		SA_1	r = 0.17, ρ = 0.04	r = -0.06, ρ = 0.55
		SA_2	r = 0.11, ρ = 0.24	r = -0.02, ρ = 0.82
		SA_3	r = 0.01, ρ = 0.9	r = 0.01, ρ = 0.92

Collective left-clicks identified targets within range of collectives, whether targets were abandoned, and was required to issue commands. The number of collective left-clicks descriptive statistics per decision difficulty are shown in Table E.6. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are shown in Tables E.7 and E.8, respectively. The Spearman correlation between the collective left-clicks per decision and selection success rate are provided in Table E.9.

Table E.6: Collective left-clicks per decision descriptive statistics by decision difficulty.

	Model	**Decision Difficulty**	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	Overall	6.53 (5.49)	5 (0/44)
		Easy	5.6 (4.53)	4 (0/26)
		Hard	7.79 (6.37)	7 (0/44)
	M_3	Overall	12.94 (7.7)	11 (2/53)
		Easy	11.77 (6.87)	10 (2/51)
		Hard	14.64 (8.51)	14 (2/53)
Collective	M_2	Overall	7.35 (5.29)	6 (0/42)
		Easy	5.96 (4.6)	5 (0/42)
		Hard	8.96 (5.59)	8 (0/27)
	M_3	Overall	12.89 (6.33)	12 (2/40)
		Easy	11.38 (5.44)	10 (2/38)
		Hard	15.04 (6.88)	14 (2/40)

Table E.7: Within model comparison statistics (DOF = 1) of collective left-clicks per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 24017, $\rho < 0.001$
	Easy	393	U = 7378.5, $\rho < 0.001$
	Hard	279	U = 4454.5, $\rho < 0.001$
Collective	Overall	672	U = 25171, $\rho < 0.001$
	Easy	377	U = 6737.5, $\rho < 0.001$
	Hard	295	U = 4952, $\rho < 0.001$

Table E.8: Between visualization comparison statistics (DOF = 1) of collective left-clicks per decision by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	672	U = 63495, $\rho < 0.01$
	Easy	374	U = 18812, $\rho = 0.19$
	Hard	298	U = 12872, $\rho = 0.02$
M_3	Overall	672	U = 58636, $\rho = 0.38$
	Easy	396	U = 19799, $\rho = 0.86$
	Hard	276	U = 10088, $\rho = 0.39$

Table E.9: Spearman correlation analysis between collective left-clicks per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = -0.11, $\rho = 0.04$	r = 0.009, $\rho = 0.87$
	Easy	r = -0.18, $\rho = 0.01$	r = -0.15, $\rho = 0.04$
	Hard	r = 0.04, $\rho = 0.67$	r = 0.23, $\rho < 0.01$
M_3	Overall	r = 0.09, $\rho = 0.11$	r = 0.04, $\rho = 0.5$
	Easy	r = 0.03, $\rho = 0.72$	r = -0.03, $\rho = 0.67$
	Hard	r = 0.26, $\rho < 0.01$	r = 0.18, $\rho = 0.03$

Intervention frequency was the total number of interventions, which occurred when the operator abandoned a target with greater than 10% support from a collective, divided by 12 decisions. Intervention frequency was assessed per participant, due to the inability to associate an intervention to a decision, and the descriptive statistics are shown in Table E.10. The Mann-Whitney-Wilcoxin within model statistical comparison

Table E.10: Intervention frequency (Number of Interventions/ Total Decisions) per participant descriptive statistics.

	Model	Mean (SD)	Median (Min/Max)
IA	M_2	0.13 (0.17)	0.04 (0/0.58)
	M_3	0.31 (0.36)	0.25 (0/1.42)
Collective	M_2	0.18 (0.17)	0.13 (0/0.58)
	M_3	0.42 (0.43)	0.29 (0/1.5)

is shown in Table E.11. No significant effects between visualizations were found.

Table E.11: Within model comparison statistics (DOF = 1) of intervention frequency (Number of Interventions/ Total Decisions) per participant.

	Sample Size	**Mann-Whitney-Wilcoxin**
IA	56	U = 270.5, ρ = 0.04
Collective	56	U = 285, ρ = 0.08

The *highlight agents* selection box identified which individual entities belonged to a particular collective. The number of times the highlight agents was used per participant, due to the inability to associate the selection of the highlight agents to a decision, and the descriptive statistics are shown in Table E.12. The highlight agents was only available on the IA interface. No significant effects between models were found.

Table E.12: Highlight agents per participant descriptive statistics.

	Model	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	0.79 (1.64)	0 (0/6)
	M_3	0.25 (0.59)	0 (0/2)

Appendix F: Additional System Design Element Usability

The sum of investigate, abandon, and decide commands was the number of operator issued *commands*. The number of commands per decision descriptive statistics are shown in Table F.1. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are shown in Tables F.2 and F.3, respectively. The Spearman correlation between the commands and selection success rate are provided in Table F.4.

Table F.1: Commands per decision descriptive statistics by decision difficulty.

	Model	**Decision Difficulty**	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	Overall	2.65 (3.49)	1 (0/31)
		Easy	2.53 (2.87)	1 (0/14)
		Hard	2.81 (4.2)	1 (0/31)
	M_3	Overall	9.93 (3.83)	10 (1/35)
		Easy	9.29 (3.99)	9 (1/35)
		Hard	10.86 (3.38)	11 (1/19)
Collective	M_2	Overall	2.43 (1.73)	2 (0/9)
		Easy	2.19 (1.64)	2 (0/9)
		Hard	2.7 (1.81)	2 (0/9)
	M_3	Overall	5.95 (2.3)	6 (1/15)
		Easy	5.42 (2.17)	5 (1/15)
		Hard	6.71 (2.26)	7 (1/13)

Table F.2: Within model comparison statistics (DOF = 1) of commands per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 8599.5, $\rho < 0.001$
	Easy	393	U = 2915, $\rho < 0.001$
	Hard	279	U = 1636.5, $\rho < 0.001$
Collective	Overall	672	U = 11903, $\rho < 0.001$
	Easy	377	U = 3755.5, $\rho < 0.001$
	Hard	295	U = 1951.5, $\rho < 0.001$

Table F.3: Between visualization comparison statistics (DOF = 1) of commands per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
M_2	Overall	672	U = 64982, $\rho < 0.001$
	Easy	374	U = 18842, $\rho = 0.18$
	Hard	298	U = 13696, $\rho < 0.001$
M_3	Overall	672	U = 18338, $\rho < 0.001$
	Easy	396	U = 6274.5, $\rho < 0.001$
	Hard	276	U = 2453.5, $\rho < 0.001$

Table F.4: Spearman correlation analysis between commands per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = 0.03, ρ = 0.59	r = 0.05, ρ = 0.35
	Easy	r = -0.009, ρ = 0.9	r = -0.02, ρ = 0.82
	Hard	r = 0.06, ρ = 0.46	r = 0.16, ρ = 0.05
M_3	Overall	r = -0.03, ρ = 0.52	r = -0.07, ρ = 0.17
	Easy	r = -0.15, ρ = 0.03	r = -0.06, ρ = 0.39
	Hard	r = 0.27, $\rho < 0.01$	r = 0.001, ρ = 0.99

Command frequency was the summation of investigate, abandon, and decide commands per decision divided by decision time. The command frequency was assessed per decision and the descriptive statistics are shown in Table F.5 [57]. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are shown in Tables F.6 and F.7, respectively. The Spearman correlation between command frequency and selection success rate are provided in Table F.8.

Table F.5: Command frequency (Number of commands/Decision time) per decision descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	0.71 (0.99)	0.37 (0/7)
		Easy	0.79 (1.03)	0.42 (0/5.53)
		Hard	0.61 (0.92)	0.31 (0/7)
	M_3	Overall	2.18 (1.45)	1.95 (0.07/13.66)
		Easy	2.24 (1.51)	1.96 (0.07/13.66)
		Hard	2.09 (1.35)	1.87 (0.1/8.36)
Collective	M_2	Overall	0.66 (0.49)	0.55 (0/2.6)
		Easy	0.7 (0.56)	0.6 (0/2.6)
		Hard	0.6 (0.4)	0.51 (0/2.03)
	M_3	Overall	1.33 (0.79)	1.18 (0.11/4.75)
		Easy	1.4 (0.87)	1.21 (0.11/4.75)
		Hard	1.22 (0.65)	1.08 (0.3/3.66)

Table F.6: Within model comparison statistics (DOF = 1) of command frequency (Number of commands/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 15640, $\rho < 0.001$
	Easy	393	U = 5847.5, $\rho < 0.001$
	Hard	279	U = 2346, $\rho < 0.001$
Collective	Overall	672	U = 23932, $\rho < 0.001$
	Easy	377	U = 8153.5, $\rho < 0.001$
	Hard	295	U = 4120.5, $\rho < 0.001$

Table F.7: Between visualization comparison statistics (DOF = 1) of command frequency (Number of commands/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
M_2	Overall	672	U = 66382, $\rho < 0.001$
	Easy	374	U = 19716, $\rho = 0.03$
	Hard	298	U = 13900, $\rho < 0.001$
M_3	Overall	672	U = 31320, $\rho < 0.001$
	Easy	396	U = 11534, $\rho < 0.001$
	Hard	276	U = 4772.5, $\rho < 0.001$

Table F.8: Spearman correlation analysis between command frequency (Number of commands/Decision time) per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = 0.08, ρ = 0.13	r = 0.1, ρ = 0.07
	Easy	r = 0.07, ρ = 0.3	r = 0.08, ρ = 0.29
	Hard	r = 0.04, ρ = 0.64	r = 0.1, ρ = 0.21
M_3	Overall	r = -0.07, ρ = 0.23	r = -0.07, ρ = 0.17
	Easy	r = -0.07, ρ = 0.35	r = -0.004, ρ = 0.96
	Hard	r = -0.12, ρ = 0.15	r = -0.19, ρ = 0.03

Investigate command frequency was the number of investigate commands issued per decision divided by decision time. The investigate command frequency was assessed per decision and the descriptive statistics are shown in Table F.9 [57]. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are shown in Tables F.10 and F.11, respectively. The Spearman correlation between investigate command frequency and selection success rate are provided in Table F.12.

Table F.9: Investigate command frequency (Number of investigate commands/Decision time) per decision descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	0.58 (0.94)	0.26 (0/6.46)
		Easy	0.65 (0.98)	0.32 (0/5.03)
		Hard	0.88 (0.48)	0.19 (0/6.46)
	M_3	Overall	1.94 (1.38)	1.7 (0/12.75)
		Easy	1.98 (1.44)	1.71 (0/12.75)
		Hard	1.88 (1.29)	1.69 (0/7.87)
Collective	M_2	Overall	0.47 (0.44)	0.35 (0/2.27)
		Easy	0.49 (0.5)	0.34 (0/2.27)
		Hard	0.46 (0.37)	0.37 (0/1.63)
	M_3	Overall	1.07 (0.71)	0.95 (0/4.3)
		Easy	1.12 (0.79)	0.97 (0/4.3)
		Hard	1 (0.58)	0.92 (0/2.81)

Table F.10: Within model comparison statistics (DOF = 1) of investigate command frequency (Number of investigate commands/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 16064, $\rho < 0.001$
	Easy	393	U = 6029.5, $\rho < 0.001$
	Hard	279	U = 2424, $\rho < 0.001$
Collective	Overall	672	U = 24483, $\rho < 0.001$
	Easy	377	U = 7957, $\rho < 0.001$
	Hard	295	U = 4379.5, $\rho < 0.001$

Table F.11: Between visualization comparison statistics (DOF = 1) of investigate command frequency (Number of investigate commands/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
M_2	Overall	672	U = 64412, $\rho < 0.01$
	Easy	374	U = 18255, $\rho = 0.44$
	Hard	298	U = 14210, $\rho < 0.001$
M_3	Overall	672	U = 29396, $\rho < 0.001$
	Easy	396	U = 10923, $\rho < 0.001$
	Hard	276	U = 4441, $\rho < 0.001$

Table F.12: Spearman correlation analysis between investigate command frequency (Number of investigate commands/Decision time) per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = 0.05, ρ = 0.34	r = 0.11, ρ = 0.04
	Easy	r = 0.08, ρ = 0.29	r = 0.08, ρ = 0.31
	Hard	r = -0.04, ρ = 0.6	r = 0.15, ρ = 0.06
M_3	Overall	r = -0.07, ρ = 0.18	r = -0.08, ρ = 0.14
	Easy	r = -0.08, ρ = 0.28	r = -0.004, ρ = 0.96
	Hard	r = -0.12, ρ = 0.18	r = -0.19, ρ = 0.02

Abandon command frequency was the number of abandon commands issued per decision divided by decision time. The abandon command frequency was assessed per decision and the descriptive statistics are shown in Table F.13 [57]. The Mann-Whitney-Wilcoxin within model statistical comparison is shown in Table F.14. No significant effects between visualizations were found. The Spearman correlation between abandon command frequency and selection success rate are provided in Table F.15.

Table F.13: Abandon command frequency (Number of abandon commands/Decision time) per decision descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	0.02 (0.07)	0 (0/0.5)
		Easy	0.01 (0.05)	0 (0/0.33)
		Hard	0.02 (0.09)	0 (0/0.5)
	M_3	Overall	0.02 (0.07)	0 (0/0.61)
		Easy	0.03 (0.08)	0 (0/0.61)
		Hard	0.02 (0.06)	0 (0/0.27)
Collective	M_2	Overall	0.02 (0.07)	0 (0/0.38)
		Easy	0.02 (0.07)	0 (0/0.38)
		Hard	0.02 (0.07)	0 (0/0.3)
	M_3	Overall	0.03 (0.07)	0 (0/0.36)
		Easy	0.03 (0.07)	0 (0/0.36)
		Hard	0.03 (0.07)	0 (0/0.35)

Table F.14: Within model comparison statistics (DOF = 1) of abandon command frequency (Number of abandon commands/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 52894, $\rho < 0.01$
	Easy	393	U = 17965, $\rho = 0.01$
	Hard	279	U = 9201, $\rho = 0.15$
Collective	Overall	672	U = 52831, $\rho = 0.01$
	Easy	377	U = 16410, $\rho = 0.02$
	Hard	295	U = 10266, $\rho = 0.19$

Table F.15: Spearman correlation analysis between abandon command frequency (Number of abandon commands/Decision time) per decision and selection success rate by decision difficulty.

	Decision Difficulty	**IA Correlation**	**Collective Correlation**
M_2	Overall	r = -0.05, $\rho = 0.37$	r = -0.02, $\rho = 0.68$
	Easy	r = -0.19, $\rho < 0.01$	r = -0.03, $\rho = 0.65$
	Hard	r = 0.09, $\rho = 0.28$	r = 0.006, $\rho = 0.94$
M_3	Overall	r = 0.09, $\rho = 0.09$	r = 0.02, $\rho = 0.76$
	Easy	r = 0.02, $\rho = 0.75$	r = 0.007, $\rho = 0.93$
	Hard	r = 0.18, $\rho = 0.03$	r = 0.03, $\rho = 0.68$

Decide command frequency was the number of decide commands issued per decision divided by decision time. The decide command frequency was assessed per decision and the descriptive statistics are shown in Table F.16 [57]. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are shown in Tables F.17 and F.18, respectively. The Spearman correlation between decide command frequency and selection success rate are provided in Table F.19.

Table F.16: Decide command frequency (Number of decide commands/Decision time) per decision descriptive statistics by decision difficulty.

	Model	**Decision Difficulty**	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	Overall	0.11 (0.15)	0 (0/0.54)
		Easy	0.12 (0.16)	0 (0/0.51)
		Hard	0.09 (0.13)	0 (0/0.54)
	M_3	Overall	0.21 (0.09)	0.2 (0/0.91)
		Easy	0.23 (0.1)	0.22 (0/0.91)
		Hard	0.18 (0.08)	0.16 (0/0.5)
Collective	M_2	Overall	0.15 (0.16)	0.14 (0/0.55)
		Easy	0.19 (0.18)	0.22 (0/0.55)
		Hard	0.1 (0.13)	0 (0/0.41)
	M_3	Overall	0.23 (0.12)	0.21 (0/1.15)
		Easy	0.26 (0.12)	0.24 (0.08/1.15)
		Hard	0.2 (0.12)	0.17 (0/0.9)

Table F.17: Within model comparison statistics (DOF = 1) of decide command frequency (Number of decide commands/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 31612, $\rho < 0.001$
	Easy	393	U = 11172, $\rho < 0.001$
	Hard	279	U = 5633.5, $\rho < 0.001$
Collective	Overall	672	U = 41268, $\rho < 0.001$
	Easy	377	U = 15076, $\rho = 0.01$
	Hard	295	U = 7013.5, $\rho < 0.001$

Table F.18: Between visualization comparison statistics (DOF = 1) of decide command frequency (Number of decide commands/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
M_2	Overall	672	U = 64685, $\rho < 0.001$
	Easy	374	U = 21377, $\rho < 0.001$
	Hard	298	U = 11598, $\rho = 0.43$
M_3	Overall	672	U = 60532, $\rho = 0.1$
	Easy	396	U = 22092, $\rho = 0.03$
	Hard	276	U = 9668.5, $\rho = 0.83$

Table F.19: Spearman correlation analysis between decide command frequency (Number of decide commands/Decision time) per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = 0.006, ρ = 0.91	r = 0.05, ρ = 0.41
	Easy	r = 0.03, ρ = 0.65	r = 0.02, ρ = 0.74
	Hard	r = -0.05, ρ = 0.52	r = -0.04, ρ = 0.65
M_3	Overall	r = -0.03, ρ = 0.56	r = -0.03, ρ = 0.53
	Easy	r = 0.07, ρ = 0.32	r = 0.03, ρ = 0.72
	Hard	r = -0.32, $\rho < 0.001$	r = -0.21, ρ = 0.01

Collective left-click frequency was the number of collective left-clicks per decision divided by decision time. The collective left-click frequency was assessed per decision and the descriptive statistics are shown in Table F.20. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are shown in Tables F.21 and F.22, respectively. The Spearman correlation between collective left-click frequency and selection success rate are provided in Table F.23.

Table F.20: Collective left-click frequency (Number of collective left-clicks/Decision time) per decision descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	1.54 (1.07)	1.41 (0/5.58)
		Easy	1.54 (1.11)	1.37 (0/5.58)
		Hard	1.54 (1.02)	1.5 (0/5.34)
	M_3	Overall	2.38 (1.12)	2.29 (0.46/7.8)
		Easy	2.4 (1.11)	2.34 (0.46/7.8)
		Hard	2.36 (1.15)	2.21 (0.63/6.66)
Collective	M_2	Overall	1.83 (1.02)	1.71 (0/4.85)
		Easy	1.79 (1.07)	1.65 (0/4.85)
		Hard	1.88 (0.96)	1.76 (0/4.26)
	M_3	Overall	2.57 (1.1)	2.45 (0.46/7.01)
		Easy	2.61 (1.13)	2.55 (0.46/7.01)
		Hard	2.5 (1.05)	2.29 (0.76/6.57)

Table F.21: Within model comparison statistics (DOF = 1) of collective left-click frequency (Number of collective left-clicks/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 31058, $\rho < 0.001$
	Easy	393	U = 10269, $\rho < 0.001$
	Hard	279	U = 5664.5, $\rho < 0.001$
Collective	Overall	672	U = 35246, $\rho < 0.001$
	Easy	377	U = 10528, $\rho < 0.001$
	Hard	295	U = 7220, $\rho < 0.001$

Table F.22: Between visualization comparison statistics (DOF = 1) of collective left-click frequency (Number of collective left-clicks/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
M_2	Overall	672	U = 66662, $\rho < 0.001$
	Easy	374	U = 20122, $\rho = 0.01$
	Hard	298	U = 13394, $\rho < 0.01$
M_3	Overall	672	U = 62655, $\rho = 0.01$
	Easy	396	U = 22072, $\rho = 0.03$
	Hard	276	U = 10450, $\rho = 0.16$

Table F.23: Spearman correlation analysis between collective left-click frequency (Number of collective left-clicks/Decision time) per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = -0.02, ρ = 0.75	r = 0.07, ρ = 0.22
	Easy	r = -0.02, ρ = 0.74	r = -0.04, ρ = 0.6
	Hard	r = -0.004, ρ = 0.96	r = 0.18, ρ = 0.03
M_3	Overall	r = 0.09, ρ = 0.1	r = -0.01, ρ = 0.79
	Easy	r = 0.11, ρ = 0.13	r = -0.03, ρ = 0.71
	Hard	r = 0.07, ρ = 0.43	r = -0.01, ρ = 0.9

Target left-click frequency was the number of target left-clicks per decision divided by decision time. The target left-click frequency was assessed per decision and the descriptive statistics are presented in Table F.24. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.25 and F.26, respectively. The Spearman correlation between target left-click frequency and selection success rate are provided in Table F.27.

Table F.24: Target left-click frequency (Number of target left-clicks/Decision time) per decision descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	1.09 (1.16)	0.74 (0/8.08)
		Easy	1.18 (1.19)	0.8 (0/5.91)
		Hard	0.97 (1.1)	0.62 (0/8.08)
	M_3	Overall	2.5 (1.57)	2.28 (0.11/13.66)
		Easy	2.55 (1.59)	2.32 (0.13/13.66)
		Hard	2.43 (1.53)	2.13 (0.11/8.36)
Collective	M_2	Overall	0.75 (0.55)	0.62 (0/3.11)
		Easy	0.79 (0.61)	0.68 (0/3.11)
		Hard	0.69 (0.48)	0.58 (0/2.43)
	M_3	Overall	1.55 (1)	1.3 (0.11/5.96)
		Easy	1.63 (1.04)	1.42 (0.11/5.23)
		Hard	1.43 (0.93)	1.2 (0.32/5.96)

Table F.25: Within model comparison statistics (DOF = 1) of target left-click frequency (Number of target left-clicks/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 20929, $\rho < 0.001$
	Easy	393	U = 7847, $\rho < 0.001$
	Hard	279	U = 3094.5, $\rho < 0.001$
Collective	Overall	672	U = 24312, $\rho < 0.001$
	Easy	377	U = 8035.5, $\rho < 0.001$
	Hard	295	U = 4368, $\rho < 0.001$

Table F.26: Between visualization comparison statistics (DOF = 1) of target left-click frequency (Number of target left-clicks/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
M_2	Overall	672	U = 51204, $\rho = 0.04$
	Easy	374	U = 15318, $\rho = 0.04$
	Hard	298	U = 10592, $\rho = 0.52$
M_3	Overall	672	U = 30752, $\rho < 0.001$
	Easy	396	U = 11360, $\rho < 0.001$
	Hard	276	U = 4641, $\rho < 0.001$

Table F.27: Spearman correlation analysis between target left-click frequency (Number of target left-clicks/Decision time) per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = 0.11, ρ = 0.04	r = 0.09, ρ = 0.11
	Easy	r = 0.13, ρ = 0.06	r = 0.1, ρ = 0.18
	Hard	r = 0.06, ρ = 0.45	r = 0.07, ρ = 0.39
M_3	Overall	r = -0.09, ρ = 0.09	r = -0.05, ρ = 0.4
	Easy	r = -0.08, ρ = 0.26	r = -0.007, ρ = 0.92
	Hard	r = -0.15, ρ = 0.08	r = -0.13, ρ = 0.14

Collective right-click frequency was the number of collective right-clicks per decision divided by decision time. The collective right-click frequency was assessed per decision and the descriptive statistics are shown in Table F.28. The collective right-click frequency was only assessed for the IA evaluation. The Mann-Whitney-Wilcoxin within model statistical comparison is shown in Table F.29. No correlations were found between collective right-click frequency and selection success rate.

Table F.28: Collective right-click frequency (Number of collective right-clicks/Decision time) per decision descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	0.36 (0.51)	0 (0/2.16)
		Easy	0.3 (0.48)	0 (0/1.92)
		Hard	0.44 (0.54)	0.21 (0/2.16)
	M_3	Overall	0.15 (0.27)	0 (0/1.37)
		Easy	0.15 (0.28)	0 (0/1.37)
		Hard	0.15 (0.26)	0 (0/1.25)

Table F.29: Within model comparison statistics (DOF = 1) of collective right-click frequency (Number of collective right-clicks/Decision time) per decision by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	672	U = 66662, $\rho < 0.001$
	Easy	393	U = 21432, $\rho = 0.03$
	Hard	279	U = 12473, $\rho < 0.001$

Target right-click frequency was the number of target right-clicks per decision divided by decision time. The target right-click frequency was assessed per decision and the descriptive statistics are presented in Table F.30. The Mann-Whitney-Wilcoxin within model statistical comparison is presented in Table F.31. No significant effects between visualizations were found. The Spearman correlation between target right-click frequency and selection success rate are provided in Table F.32.

Table F.30: Target right-click frequency (Number of target right-clicks/Decision time) per decision descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	0.83 (0.94)	0.51 (0/5.22)
		Easy	0.71 (0.83)	0.45 (0/4.72)
		Hard	1 (1.06)	0.64 (0/5.22)
	M_3	Overall	0.66 (0.7)	0.41 (0/4.76)
		Easy	0.72 (0.77)	0.45 (0/4.76)
		Hard	0.58 (0.58)	0.33 (0/3.3)
Collective	M_2	Overall	0.77 (0.85)	0.54 (0/5.55)
		Easy	0.77 (0.84)	0.54 (0/4.69)
		Hard	0.76 (0.88)	0.55 (0/5.55)
	M_3	Overall	0.59 (0.62)	0.36 (0/3.27)
		Easy	0.65 (0.66)	0.41 (0/3.11)
		Hard	0.51 (0.55)	0.3 (0/3.27)

Table F.31: Within model comparison statistics (DOF = 1) of target right-click frequency (Number of target right-clicks/Decision time) per decision by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	672	U = 58750, ρ = 0.36
	Easy	393	U = 18048, ρ = 0.26
	Hard	279	U = 11632, $\rho < 0.01$
Collective	Overall	672	U = 63119, $\rho < 0.01$
	Easy	377	U = 19244, ρ = 0.15
	Hard	295	U = 12798, $\rho < 0.01$

Table F.32: Spearman correlation analysis between target right-click frequency (Number of target right-clicks/Decision time) per decision and selection success rate by decision difficulty.

	Decision Difficulty	**IA Correlation**	**Collective Correlation**
M_2	Overall	r = -0.05, ρ = 0.33	r = 0.12, ρ = 0.03
	Easy	r = -0.003, ρ = 0.96	r = 0.09, ρ = 0.25
	Hard	r = -0.07, ρ = 0.44	r = 0.14, ρ = 0.09
M_3	Overall	r = 0.09, ρ = 0.09	r = 0.03, ρ = 0.56
	Easy	r = 0.04, ρ = 0.56	r = 0.07, ρ = 0.3
	Hard	r = 0.05, ρ = 0.6	r = -0.04, ρ = 0.62

Collective and target left- and right-clicks were examined per participant. *Target left-clicks* were the second click required in the process of issuing commands, but did not provide supplementary information. The number of collective and target left- and right-clicks descriptive statistics are presented in Table F.33 [177]. No significant effects between visualizations were found.

Table F.33: Collective and target left- and right-clicks per participant descriptive statistics.

	Clicks	Mean (SD)	Median (Min/Max)
IA	Collective Left	107.6 (49.89)	104 (5/235)
	Collective Right	30.64 (20.98)	27.5 (0/85)
	Target Left	97.64 (58.78)	83 (5/251)
	Target Right	97.18 (82.79)	68.5 (4/352)
Collective	Collective Left	121.96 (47.4)	130.5 (35/212)
	Collective Right	30.57 (31.95)	19.5 (7/164)
	Target Left	185.6 (64.32)	202 (62/290)
	Target Right	82.39 (60.22)	75 (23/278)

The *cancel abandon command* enabled an operator to cancel a previously issued abandon command for a particular collective and target. Cancel abandon commands were assessed per participant, due to the inability to associate a cancel abandon command to a decision, and the descriptive statistics are shown in Table F.34. No significant effects between models or visualizations were found.

Table F.34: Cancel abandon commands per participant descriptive statistics.

	Model	Mean (SD)	Median (Min/Max)
IA	M_2	0.68 (1.83)	0 (0/9)
	M_3	0.46 (1.71)	0 (0/9)
Collective	M_2	0.36 (0.99)	0 (0/4)
	M_3	0.71 (1.38)	0 (0/5)

The *total number of abandon commands* issued per participant was assessed and the descriptive statistics are presented in Table F.35. The Mann-Whitney-Wilcoxin within model statistical comparison is presented in Table F.14. No significant effects between visualizations were found.

Table F.35: Total number of abandon commands per participant descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	9.75 (13.4)	5 (0/59)
		Easy	3.79 (5.09)	1.5 (0/16)
		Hard	5.96 (9.66)	2.5 (0/43)
	M_3	Overall	36.79 (24.12)	42 (0/76)
		Easy	17.07 (12.45)	19 (0/34)
		Hard	19.71 (12.48)	21 (0/44)
Collective	M_2	Overall	12.04 (14)	7 (0/52)
		Easy	5.07 (7.17)	2 (0/28)
		Hard	6.96 (7.47)	4.5 (0/26)
	M_3	Overall	30.39 (24.37)	30.5 (0/68)
		Easy	15.86 (12.58)	19 (0/35)
		Hard	14.54 (12.71)	15 (0/33)

Table F.36: Within model comparison statistics (DOF = 1) of the total number of abandon commands per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	56	U = 162.5, $\rho < 0.001$
	Easy	56	U = 161.5, $\rho < 0.001$
	Hard	56	U = 171, $\rho < 0.001$
Collective	Overall	56	U = 246.5, $\rho = 0.02$
	Easy	56	U = 220.5, $\rho < 0.01$
	Hard	56	U = 291.5, $\rho = 0.1$

The *average number of abandon commands* issued per participant was also assessed and the descriptive statistics are presented in Table F.37. The Mann-Whitney-Wilcoxin within model statistical comparison is presented in Table F.38. No significant effects between visualizations were found.

Table F.37: Average number of abandon commands per participant descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	1.86 (0.98)	1.54 (1/5.03)
		Easy	1.73 (0.57)	1.6 (1/2.69)
		Hard	1.95 (1.23)	1.55 (1/5.91)
	M_3	Overall	3.45 (1.34)	3.66 (1/6.36)
		Easy	3.29 (1.25)	3.47 (1/5.56)
		Hard	3.75 (1.33)	3.72 (1.67/7.52)
Collective	M_2	Overall	1.76 (0.91)	1.36 (1/4.13)
		Easy	1.68 (0.97)	1.13 (1/4.25)
		Hard	1.8 (0.95)	1.43 (1/4)
	M_3	Overall	3.13 (1.29)	3.55 (1/4.91)
		Easy	3.3 (1.3)	3.74 (1/5.06)
		Hard	3.01 (1.3)	3.53 (1/4.76)

Table F.38: Within model comparison statistics (DOF = 1) of the average number of abandon commands per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	47	U = 89, $\rho < 0.001$
	Easy	38	U = 48, $\rho < 0.001$
	Hard	43	U = 56.5, $\rho < 0.001$
Collective	Overall	49	U = 126.5, $\rho < 0.001$
	Easy	42	U = 72, $\rho < 0.001$
	Hard	47	U = 137, $\rho < 0.01$

The average number of *targets in range when an abandon command was issued* per participant was assessed and the descriptive statistics are presented in Table F.39. The Mann-Whitney-Wilcoxin within model statistical comparison is presented in Table F.40. No significant effects between visualizations were found.

Table F.39: Targets in range when abandon command issued per participant descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	5.79 (0.85)	5.85 (4/8)
		Easy	5.35 (0.95)	5.33 (3/6.71)
		Hard	5.97 (0.99)	6 (4/8)
	M_3	Overall	5.49 (0.84)	5.64 (3/6.4)
		Easy	5.65 (1.03)	5.85 (3/6.75)
		Hard	5.52 (0.48)	5.47 (4.8/6.31)
Collective	M_2	Overall	5.57 (0.66)	5.72 (4.45/6.67)
		Easy	5.51 (1.11)	5.56 (3.75/7.67)
		Hard	5.67 (0.77)	5.67 (4/7)
	M_3	Overall	5.79 (0.58)	5.9 (4/7)
		Easy	5.91 (0.59)	6.02 (4/6.71)
		Hard	5.68 (0.64)	5.65 (4.64/7)

Table F.40: Within model comparison statistics (DOF = 1) of the targets in range when abandon command issued per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	47	U = 313.5, ρ = 0.42
	Easy	38	U = 126.5, ρ = 0.17
	Hard	43	U = 330, ρ = 0.02
Collective	Overall	49	U = 233, ρ = 0.18
	Easy	42	U = 162.5, ρ = 0.15
	Hard	47	U = 278, ρ = 0.96

The average number of *targets that were abandoned when an abandon command was issued* per participant was assessed and the descriptive statistics are presented in Table F.41. The Mann-Whitney-Wilcoxin within model statistical comparison is presented in Table F.42. No significant effects between visualizations were found.

Table F.41: Abandoned targets when abandon command issued per participant descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	1.69 (0.69)	1.5 (1/3.2)
		Easy	1.63 (0.51)	1.57 (1/2.69)
		Hard	1.76 (0.81)	1.5 (1/3.47)
	M_3	Overall	2.75 (0.81)	2.93 (1/4)
		Easy	2.62 (0.81)	2.96 (1/3.94)
		Hard	3.03 (0.74)	3.14 (1.67/4.07)
Collective	M_2	Overall	1.61 (0.71)	1.33 (1/3.29)
		Easy	1.53 (0.75)	1.13 (1/3.46)
		Hard	1.64 (0.73)	1.33 (1/3.3)
	M_3	Overall	2.66 (0.97)	2.98 (1/4.04)
		Easy	2.71 (0.94)	2.94 (1/4.21)
		Hard	2.66 (1.04)	3.05 (1/3.91)

Table F.42: Within model comparison statistics (DOF = 1) of abandoned targets when abandon command issued per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	47	U = 95, $\rho < 0.001$
	Easy	38	U = 50.5, $\rho < 0.001$
	Hard	43	U = 61.5, $\rho < 0.001$
Collective	Overall	49	U = 117, $\rho < 0.001$
	Easy	42	U = 74, $\rho < 0.001$
	Hard	47	U = 122, $\rho < 0.001$

The average number of *individual collective entities that were favoring a target when an abandon command was issued* per participant was assessed and the descriptive statistics are shown in Table F.43. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.44 and F.45, respectively.

Table F.43: Number of favoring individual collective entities when an abandon command issued per participant descriptive statistics by decision difficulty.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	20.15 (10.8)	19.62 (1.67/46.86)
		Easy	20.5 (16.35)	14.5 (1.67/58.33)
		Hard	23.04 (13.29)	20 (6.5/49)
	M_3	Overall	12.3 (3.85)	11.97 (6.09/22.76)
		Easy	11.17 (3.95)	11.16 (5.35/22.83)
		Hard	12.73 (4.2)	12.95 (6.81/22.71)
Collective	M_2	Overall	28.04 (18.56)	23 (7.43/84.5)
		Easy	25.66 (25.3)	15.7 (6.13/82)
		Hard	31.35 (22.25)	27 (5/87)
	M_3	Overall	16.56 (7.61)	14.63 (7.86/44.5)
		Easy	13.13 (4.25)	13.06 (6.77/20.78)
		Hard	18.55 (7.63)	17.04 (9.79/44.5)

Table F.44: Within model comparison statistics (DOF = 1) of the number of favoring individual collective entities when an abandon command issued per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	47	U = 413, $\rho < 0.01$
	Easy	38	U = 218, $\rho = 0.18$
	Hard	43	U = 349, $\rho < 0.01$
Collective	Overall	49	U = 402, $\rho = 0.04$
	Easy	42	U = 248, $\rho = 0.49$
	Hard	47	U = 363, $\rho = 0.06$

Table F.45: Between visualization comparison statistics (DOF = 1) of the number of favoring individual collective entities when an abandon command issued per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	47	U = 327.5, ρ = 0.27
	Easy	35	U = 153, ρ = 0.93
	Hard	45	U = 300, ρ = 0.26
M_3	Overall	49	U = 426, ρ = 0.01
	Easy	45	U = 315, ρ = 0.16
	Hard	45	U = 393, $\rho < 0.01$

The *average number of decide commands* issued per participant was also assessed and the descriptive statistics are presented in Table F.46. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparison are presented in Tables F.47 and F.48, respectively.

Table F.46: Average number of decide commands per participant descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	4.96 (3.51)	5.5 (0/11)
		Easy	2.36 (1.93)	2 (0/6)
		Hard	2.54 (2.1)	2 (0/7)
	M_3	Overall	13.36 (1.16)	13 (12/15)
		Easy	6.54 (0.74)	6 (6/8)
		Hard	6.82 (0.77)	7 (6/8)
Collective	M_2	Overall	6.36 (4)	6 (0/13)
		Easy	3.68 (2.45)	3.5 (0/7)
		Hard	2.68 (1.83)	3 (0/6)
	M_3	Overall	12.82 (1.56)	13 (7/16)
		Easy	6.64 (0.87)	6 (6/8)
		Hard	6.18 (1.22)	6 (1/8)

Table F.47: Within model comparison statistics (DOF = 1) of the average number of decide commands per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	56	U = 0, $\rho < 0.001$
	Easy	56	U = 17, $\rho < 0.001$
	Hard	56	U = 22, $\rho < 0.001$
Collective	Overall	56	U = 38.5, $\rho < 0.001$
	Easy	56	U = 110, $\rho < 0.001$
	Hard	56	U = 31.5, $\rho < 0.001$

Table F.48: Between visualization comparison statistics (DOF = 1) of the average number of decide commands per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	56	U = 472, $\rho = 0.19$
	Easy	56	U = 521, $\rho = 0.03$
	Hard	56	U = 414, $\rho = 0.72$
M_3	Overall	56	U = 316, $\rho = 0.2$
	Easy	56	U = 408.5, $\rho = 0.76$
	Hard	56	U = 251, $\rho = 0.01$

The average number of *individual collective entities that were favoring a target when the collective was in the commit state* per participant was assessed and the descriptive statistics are shown in Table F.49. The Mann-Whitney-Wilcoxin within model statistical comparisons are presented in Table F.50. No significant effects between visualizations were found.

Table F.49: Average number of favoring individual collective entities in committed state per participant descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	60.48 (0.41)	60.33 (60/61.55)
		Easy	60.46 (0.38)	60.5 (60/61.17)
		Hard	60.49 (0.57)	60.37 (60/62)
	M_3	Overall	60.65 (0.4)	60.58 (60.07/61.75)
		Easy	60.5 (0.35)	60.41 (60/61.43)
		Hard	60.78 (0.61)	60.71 (60/62.33)
Collective	M_2	Overall	60.71 (0.43)	60.69 (60/61.38)
		Easy	60.69 (0.53)	60.6 (60/62)
		Hard	60.77 (0.69)	60.6 (60/62.5)
	M_3	Overall	60.82 (0.88)	60.63 (60/64.62)
		Easy	60.58 (0.47)	60.5 (60/61.83)
		Hard	61.11 (1.65)	60.7 (60/68.67)

Table F.50: Within model comparison statistics (DOF = 1) of the average number of favoring individual collective entities in committed state per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	51	U = 241.5, ρ = 0.13
	Easy	48	U = 264.5, ρ = 0.75
	Hard	50	U = 199.5, ρ = 0.03
Collective	Overall	52	U = 359, ρ = 0.68
	Easy	51	U = 362.5, ρ = 0.45
	Hard	51	U = 301, ρ = 0.7

The average number of *individual collective entities that were favoring a target when a decide command was issued* per participant was assessed and the descriptive statistics are shown in Table F.51. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.52 and F.53, respectively.

Table F.51: Average number of favoring individual collective entities when a decide command issued per participant descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	95.75 (7.81)	95 (82/112.5)
		Easy	91.62 (7.47)	90.75 (77.5/107.75)
		Hard	99.79 (11.06)	98.04 (81/127)
	M_3	Overall	69.2 (5.52)	67.71 (64.14/85.58)
		Easy	67.37 (4.99)	66.17 (62.33/86.5)
		Hard	70.98 (7.68)	68.77 (63.5/98.5)
Collective	M_2	Overall	96.3 (7.27)	96.5 (85/119.67)
		Easy	90.88 (11.37)	92.67 (68.8/126)
		Hard	103.38 (9.49)	105.6 (75/116.5)
	M_3	Overall	72.43 (5.34)	70.54 (65.29/85.93)
		Easy	69.63 (5.01)	68.33 (63.17/84)
		Hard	75.32 (7.69)	73.25 (64.83/93.67)

Table F.52: Within model comparison statistics (DOF = 1) of the average number of favoring individual collective entities when a decide command issued per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	51	U = 639.5, $\rho < 0.001$
	Easy	48	U = 554, $\rho < 0.001$
	Hard	50	U = 600, $\rho < 0.001$
Collective	Overall	52	U = 671, $\rho < 0.001$
	Easy	51	U = 620, $\rho < 0.001$
	Hard	51	U = 626, $\rho < 0.001$

Table F.53: Between visualization comparison statistics (DOF = 1) of the average number of favoring individual collective entities when a decide command issued per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	47	U = 288.5, ρ = 0.8
	Easy	43	U = 227, ρ = 0.95
	Hard	45	U = 331.5, ρ = 0.08
M_3	Overall	56	U = 576, $\rho < 0.01$
	Easy	56	U = 518.5, ρ = 0.04
	Hard	56	U = 568, $\rho < 0.01$

The average number of *individual collective entities that were committed to a target when the collective begins executing* per participant was assessed and the descriptive statistics are shown in Table F.54. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.55 and F.56, respectively.

Table F.54: Average number of committed individual collective entities when collective begins executing per participant descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	75.51 (16.66)	79 (32/107)
		Easy	87.18 (21.28)	88.13 (49/121.33)
		Hard	63.97 (16.31)	65.04 (32/89)
	M_3	Overall	38.62 (5.21)	37.95 (22.6/49.5)
		Easy	41.56 (7.91)	41.42 (19.13/54.67)
		Hard	35.74 (7.57)	35.48 (19.29/57.33)
Collective	M_2	Overall	82.66 (8.21)	83.03 (60.88/94.88)
		Easy	97.74 (13.74)	97 (80/137)
		Hard	65.36 (9.37)	64.8 (45.2/80)
	M_3	Overall	41.4 (5.07)	40.58 (30.79/51.17)
		Easy	45.02 (7.43)	47.04 (24.17/56)
		Hard	37.55 (8.37)	33.65 (24.67/52.17)

Table F.55: Within model comparison statistics (DOF = 1) of the average number of committed individual collective entities when collective begins executing per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	51	U = 616, $\rho < 0.001$
	Easy	48	U = 552, $\rho < 0.001$
	Hard	50	U = 574, $\rho < 0.001$
Collective	Overall	52	U = 672, $\rho < 0.001$
	Easy	51	U = 644, $\rho < 0.001$
	Hard	51	U = 636, $\rho < 0.001$

Table F.56: Between visualization comparison statistics (DOF = 1) of the average number of committed individual collective entities when collective begins executing per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	47	U = 356, $\rho = 0.09$
	Easy	43	U = 286, $\rho = 0.18$
	Hard	45	U = 261, $\rho = 0.86$
M_3	Overall	56	U = 511, $\rho = 0.05$
	Easy	56	U = 499, $\rho = 0.08$
	Hard	56	U = 426, $\rho = 0.58$

The average number of *individual collective entities that were executing when the collective begins executing* per participant was assessed and the descriptive statistics are shown in Table F.57. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.58 and F.59, respectively.

Table F.57: Average number of executing individual collective entities when collective begins executing per participant descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	15.18 (10.65)	14.33 (1/52)
		Easy	11.38 (9.03)	8 (2.33/35.5)
		Hard	19.54 (11.78)	18.72 (1/52)
	M_3	Overall	22.65 (4.54)	22.16 (15.25/33.2)
		Easy	25.34 (8.43)	25.67 (12.17/44.75)
		Hard	20.19 (7.02)	20 (5/35)
Collective	M_2	Overall	9.74 (4.14)	9.78 (1.5/21.5)
		Easy	8.3 (3.85)	8.17 (2/14.71)
		Hard	10.86 (7.29)	11 (1/27.2)
	M_3	Overall	19.11 (4.98)	19.23 (10.21/29.08)
		Easy	20.54 (9.55)	19.08 (7.75/49.5)
		Hard	17.74 (8.18)	18.47 (3.5/35)

Table F.58: Within model comparison statistics (DOF = 1) of the average number of executing individual collective entities when collective begins executing per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	51	U = 97, $\rho < 0.001$
	Easy	48	U = 67, $\rho < 0.001$
	Hard	50	U = 277.5, $\rho = 0.56$
Collective	Overall	52	U = 51, $\rho < 0.001$
	Easy	51	U = 41, $\rho < 0.001$
	Hard	51	U = 168, $\rho < 0.01$

Table F.59: Between visualization comparison statistics (DOF = 1) of the average number of executing individual collective entities when collective begins executing per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	47	U = 162.5, ρ = 0.02
	Easy	43	U = 207, ρ = 0.58
	Hard	45	U = 133, $\rho < 0.01$
M_3	Overall	56	U = 237.5, ρ = 0.01
	Easy	56	U = 246.5, ρ = 0.02
	Hard	56	U = 315.5, ρ = 0.21

The *time difference (minutes) between an issued decide command and executing collective* per participant was assessed and the descriptive statistics are shown in Table F.60. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.61 and F.62, respectively.

Table F.60: Time (minutes) between issued decide command and executing collective per participant descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	0.1 (0.01)	0.1 (0.08/0.12)
		Easy	0.11 (0.01)	0.11 (0.08/0.14)
		Hard	0.09 (0.01)	0.09 (0.07/0.13)
	M_3	Overall	0.09 (0.01)	0.09 (0.07/0.11)
		Easy	0.09 (0.01)	0.09 (0.07/0.11)
		Hard	0.1 (0.01)	0.1 (0.07/0.12)
Collective	M_2	Overall	0.1 (0.01)	0.1 (0.07/0.11)
		Easy	0.11 (0.01)	0.11 (0.08/0.13)
		Hard	0.08 (0.01)	0.08 (0.03/0.1)
	M_3	Overall	0.09 (0.01)	0.09 (0.07/0.1)
		Easy	0.08 (0.01)	0.09 (0.07/0.1)
		Hard	0.09 (0.01)	0.09 (0.07/0.11)

Table F.61: Within model comparison statistics (DOF = 1) of the time (minutes) between issued decide command and executing collective per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	51	U = 458.5, $\rho < 0.01$
	Easy	48	U = 520, $\rho < 0.001$
	Hard	50	U = 221, $\rho = 0.08$
Collective	Overall	52	U = 461, $\rho = 0.02$
	Easy	51	U = 615, $\rho < 0.001$
	Hard	51	U = 145, $\rho < 0.001$

Table F.62: Between visualization comparison statistics (DOF = 1) of the time (minutes) between issued decide command and executing collective per participant by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	47	U = 203.5, $\rho = 0.11$
	Easy	43	U = 219.5, $\rho = 0.8$
	Hard	45	U = 127.5, $\rho < 0.01$
M_3	Overall	56	U = 304, $\rho = 0.13$
	Easy	56	U = 356.5, $\rho = 0.55$
	Hard	56	U = 286.5, $\rho = 0.08$

Further analysis of how operators used the collective and target information pop-up windows was conducted. The *number of targets in range per decision* was assessed and the descriptive statistics are shown in Table F.63. No significant effects within models and between visualizations were found. The Spearman correlation between the number of targets in range per decision and selection success rate are provided in Table F.64.

Table F.63: Number of targets in range per decision descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	5.88 (1.27)	6 (2/9)
		Easy	5.81 (1.32)	6 (2/9)
		Hard	5.97 (1.18)	6 (3/9)
	M_3	Overall	5.74 (1.2)	6 (2/9)
		Easy	5.62 (1.33)	6 (2/9)
		Hard	5.9 (0.98)	6 (3/8)
Collective	M_2	Overall	5.82 (1.23)	6 (2/9)
		Easy	5.74 (1.26)	6 (2/9)
		Hard	5.92 (1.2)	6 (2/9)
	M_3	Overall	5.76 (1.19)	6 (3/9)
		Easy	5.73 (1.24)	6 (3/9)
		Hard	5.81 (1.13)	6 (3/9)

Table F.64: Spearman correlation analysis between the number of targets in range per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = 0.06, ρ = 0.26	r = 0.09, ρ = 0.09
	Easy	r = 0.1, ρ = 0.17	r = 0.29, $\rho < 0.001$
	Hard	r = 0.04, ρ = 0.66	r = -0.007, ρ = 0.93
M_3	Overall	r = 0.09, ρ = 0.08	r = -0.007, ρ = 0.9
	Easy	r = 0.14, ρ = 0.04	r = 0.04, ρ = 0.59
	Hard	r = 0.11, ρ = 0.18	r = -0.04, ρ = 0.62

The *number of targets in range with open information pop-up windows per decision* was assessed and the descriptive statistics are shown in Table F.65. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.66 and F.67, respectively. The Spearman correlation between the number of targets in range with open information pop-up windows per decision and selection success rate are provided in Table F.68.

Table F.65: Number of targets in range with open information pop-up windows per decision descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	2.41 (1.62)	2 (0/7)
		Easy	2.81 (1.67)	3 (0/7)
		Hard	2.46 (1.39)	2 (0/6)
	M_3	Overall	2.12 (1.52)	2 (0/7)
		Easy	2.57 (1.36)	2 (0/6)
		Hard	2.74 (1.31)	3 (1/6)
Collective	M_2	Overall	2.67 (1.67)	2 (0/8)
		Easy	2.92 (1.7)	3 (0/8)
		Hard	2.75 (1.71)	2 (0/8)
	M_3	Overall	2.46 (1.61)	2 (0/6)
		Easy	2.74 (1.68)	2 (0/8)
		Hard	2.74 (1.64)	3 (0/7)

Table F.66: Within model comparison statistics (DOF = 1) of the number of targets in range with open information pop-up windows per decision by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	672	U = 52269, ρ = 0.09
	Easy	393	U = 16606, ρ = 0.01
	Hard	297	U = 9905.5, ρ = 0.79
Collective	Overall	672	U = 55354, ρ = 0.66
	Easy	377	U = 16118, ρ = 0.12
	Hard	295	U = 11520, ρ = 0.35

Table F.67: Between visualization comparison statistics (DOF = 1) of the number of targets in range with open information pop-up windows per decision by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	672	U = 61188, ρ = 0.06
	Easy	374	U = 19470, ρ = 0.05
	Hard	298	U = 11354, ρ = 0.7
M_3	Overall	672	U = 58390, ρ = 0.43
	Easy	396	U = 21166, ρ = 0.16
	Hard	276	U = 9245.5, ρ = 0.67

Table F.68: Spearman correlation analysis between the number of targets in range with open information pop-up windows per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = -0.22, $\rho < 0.001$	r = -0.03, ρ = 0.6
	Easy	r = -0.23, $\rho < 0.01$	r = -0.15, ρ = 0.05
	Hard	r = -0.16, ρ = 0.05	r = 0.11, ρ = 0.19
M_3	Overall	r = -0.25, $\rho < 0.001$	r = -0.1, ρ = 0.07
	Easy	r = -0.25, $\rho < 0.001$	r = -0.07, ρ = 0.32
	Hard	r = -0.22, $\rho < 0.01$	r = -0.13, ρ = 0.13

The maximum number of times target information pop-up windows were opened per target per decision was assessed and the descriptive statistics are shown in Table F.69. The Mann-Whitney-Wilcoxin within model and between visualization statistical comparisons are presented in Tables F.70 and F.71, respectively. The Spearman correlation between the maximum number of times target information pop-up windows were opened per target per decision and selection success rate are provided in Table F.72.

Table F.69: Maximum number of times target information pop-up windows opened per target per decision descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	2.87 (2.46)	2 (1/14)
		Easy	3.57 (2.85)	3 (1/14)
		Hard	2.55 (2.88)	2 (1/30)
	M_3	Overall	2.33 (1.95)	2 (1/14)
		Easy	2.56 (2.62)	2 (1/30)
		Hard	2.58 (2.21)	2 (1/13)
Collective	M_2	Overall	2.37 (2)	2 (1/17)
		Easy	2.68 (2.46)	2 (1/17)
		Hard	2.26 (1.72)	2 (1/10)
	M_3	Overall	2.1 (1.43)	2 (1/8)
		Easy	2.31 (1.78)	2 (1/10)
		Hard	2.39 (1.85)	2 (1/10)

Table F.70: Within model comparison statistics (DOF = 1) of the maximum number of times target information pop-up windows opened per target per decision by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	637	U = 54596, ρ = 0.07
	Easy	368	U = 16492, ρ = 0.72
	Hard	269	U = 11010, $\rho < 0.01$
Collective	Overall	643	U = 52284, ρ = 0.78
	Easy	358	U = 15353, ρ = 0.52
	Hard	285	U = 10751, ρ = 0.35

Table F.71: Between visualization comparison statistics (DOF = 1) of the maximum number of times target information pop-up windows opened per target per decision by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
M_2	Overall	619	U = 43318, ρ = 0.03
	Easy	337	U = 13864, ρ = 0.69
	Hard	282	U = 7901.5, $\rho < 0.01$
M_3	Overall	661	U = 53181, ρ = 0.54
	Easy	389	U = 18786, ρ = 0.91
	Hard	272	U = 8761, ρ = 0.43

Table F.72: Spearman correlation analysis between the maximum number of times target information pop-up windows opened per target per decision and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = -0.25, $\rho < 0.001$	r = -0.03, ρ = 0.6
	Easy	r = -0.21, $\rho < 0.01$	r = -0.12, ρ = 0.11
	Hard	r = -0.19, ρ = 0.03	r = 0.06, ρ = 0.43
M_3	Overall	r = -0.05, ρ = 0.35	r = 0.07, ρ = 0.21
	Easy	r = -0.1, ρ = 0.15	r = 0.11, ρ = 0.12
	Hard	r = 0.04, ρ = 0.63	r = 0.04, ρ = 0.61

The *maximum percentage of time a target information pop-up window was open per target* relative to the decision time mean was assessed and the descriptive statistics are shown in Table F.73. The Mann-Whitney-Wilcoxin within model statistical comparisons are presented in Table F.74. No significant effects between visualizations were found. The Spearman correlation between the maximum percentage of time a target information pop-up window was open per target relative to the decision time and selection success rate are provided in Table F.75.

Table F.73: Maximum time target information pop-up windows opened per target per decision (%) descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	36.2 (34.56)	20.68 (0.17/100)
		Easy	31.73 (31.41)	15.72 (0.17/100)
		Hard	53.13 (29.18)	59.78 (0.52/100)
	M_3	Overall	39.67 (36.53)	22.89 (0.49/100)
		Easy	52.75 (29.26)	57.8 (0.52/100)
		Hard	52.2 (29.46)	55.38 (1.05/100)
Collective	M_2	Overall	40.02 (34.89)	34.26 (0.22/100)
		Easy	37.38 (33.89)	28.88 (0.34/100)
		Hard	50.03 (33.28)	52.3 (0.18/100)
	M_3	Overall	42.38 (35.7)	38.13 (0.22/100)
		Easy	49.48 (31.98)	52.54 (0.18/100)
		Hard	48.7 (30.17)	52.64 (0.44/100)

Table F.74: Within model comparison statistics (DOF = 1) of the maximum time target information pop-up windows opened per target per decision (%) by decision difficulty.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	Overall	637	U = 36398, $\rho < 0.001$
	Easy	368	U = 13208, $\rho < 0.001$
	Hard	269	U = 5774, $\rho < 0.001$
Collective	Overall	643	U = 43800, $\rho < 0.001$
	Easy	358	U = 14052, $\rho = 0.05$
	Hard	285	U = 8154, $\rho < 0.01$

Table F.75: Spearman correlation analysis between the maximum time target information pop-up windows opened per target per decision (%) and selection success rate by decision difficulty.

	Decision Difficulty	IA Correlation	Collective Correlation
M_2	Overall	r = 0.04, ρ = 0.49	r = -0.008, ρ = 0.89
	Easy	r = -0.04, ρ = 0.63	r = -0.07, ρ = 0.39
	Hard	r = 0.1, ρ = 0.27	r = 0.007, ρ = 0.93
M_3	Overall	r = -0.11, ρ = 0.04	r = -0.06, ρ = 0.3
	Easy	r = -0.07, ρ = 0.34	r = -0.08, ρ = 0.29
	Hard	r = -0.17, ρ = 0.05	r = -0.05, ρ = 0.56

Appendix G: Additional System Design Element Influence on Team Performance

The first twelve decisions were assessed for each trial and the descriptive statistics are shown in Table G.1. The Mann-Whitney-Wilcoxin within model statistical comparisons are presented in Table G.2. No significant effects between visualizations were found.

Table G.1: Number of decisions per participant by decision difficulty descriptive statistics.

	Model	**Decision Difficulty**	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	Overall	12 (0)	12 (12/12)
		Easy	6.93 (0.77)	7 (6/8)
		Hard	5.07 (0.77)	5 (4/6)
	M_3	Overall	12 (0)	12 (12/12)
		Easy	7.11 (1.03)	7 (5/9)
		Hard	4.89 (1.03)	5 (3/7)
Collective	M_2	Overall	12 (0)	12 (12/12)
		Easy	6.43 (0.96)	7 (5/8)
		Hard	5.57 (0.96)	5 (4/7)
	M_3	Overall	12 (0)	12 (12/12)
		Easy	7.04 (0.96)	7 (6/9)
		Hard	4.96 (0.96)	5 (3/6)

Table G.2: Within model comparison statistics (DOF = 1) of the number of decisions per participant by decision difficulty.

	Decision Difficulty	**Sample Size**	**Mann-Whitney-Wilcoxin**
IA	Overall	56	U = 507.5, ρ = 0.05
	Easy	56	U = 330, ρ = 0.3
	Hard	56	U = 581, $\rho < 0.01$
Collective	Overall	56	U = 445.5, ρ = 0.35
	Easy	56	U = 412.5, ρ = 0.74
	Hard	56	U = 459.5, ρ = 0.26

The decision time improvement of the human-collective team using the M_2 model over the M_{2SIM} model was assessed and the descriptive statistics are presented in Table G.3. No significant effects between visualizations were found.

Table G.3: Decision time improvement (%) of human-collective team over model descriptive statistics.

	Model	**Decision Difficulty**	**Mean (SD)**	**Median (Min/Max)**
IA	M_2	Overall	-9.91 (15.41)	-11.47 (-34.26/31.91)
		Easy	-8.92 (24.52)	-11.88 (-46.25/77.86)
		Hard	-10.76 (16.34)	-15.8 (-33.53/35.63)
Collective	M_2	Overall	-16.71 (10.9)	-17.64 (-37.36/5.06)
		Easy	-19.19 (14.34)	-21.67 (-43.04/22.82)
		Hard	-18.96 (10.35)	-18.27 (-39.12/-0.8)

The selection success rate improvement of the human-collective team using the M_2 model over the M_{2SIM} model was assessed and the descriptive statistics are presented in Table G.4. The Mann-Whitney-Wilcoxin between visualization statistical comparisons are presented in Table G.5.

Table G.4: Success rate improvement (%) of human-collective team over model descriptive statistics.

	Model	Decision Difficulty	Mean (SD)	Median (Min/Max)
IA	M_2	Overall	2.33 (24.99)	0.59 (-53.49/46.67)
		Easy	5.15 (30.91)	11.62 (-77.83/55.42)
		Hard	11.79 (52.2)	10.59 (-100/151.94)
Collective	M_2	Overall	18.86 (12.8)	19.57 (-6.25/39.53)
		Easy	24.38 (14.8)	24.14 (-11.55/59.81)
		Hard	29.1 (35.37)	26.22 (-40.55/120.86)

Table G.5: Between visualization comparison statistics (DOF = 1) of the success rate improvement (%) of human-collective team over model by decision difficulty.

Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
Overall	56	U = 544.5, $\rho = 0.01$
Easy	56	U = 560, $\rho < 0.01$
Hard	56	U = 492, $\rho = 0.1$

The SA probe improvement of Trial 1 over Trial 2 using both models and visualizations were assessed and the descriptive statistics are presented in Table G.6.

Table G.6: SA probe improvement (%) of Trial 1 over Trial 2 descriptive statistics.

	Model	Mean (SD)	Median (Min/Max)
IA	M_2	12.32 (42.37)	0 (-66.67/150)
	M_3	2.47 (53.33)	-16.67 (-66.67/200)
Collective	M_2	-3.1 (23.31)	0 (-66.67/66.67)
	M_3	25.18 (37.99)	20 (-25/100)

The *time (minutes) to respond to a SA probe question* was assessed only for the IA evaluation, because the Collective evaluation did not record response times, and the descriptive statistics are shown in Table G.7. The Mann-Whitney-Wilcoxin within model statistical comparisons are presented in Table G.8. The Spearman correlation between the SA probe response time and selection success rate are provided in Table G.9.

Table G.7: SA probe response time (minutes) descriptive statistics by SA level.

	Model	SA Level	Mean (SD)	Median (Min/Max)
IA	M_2	SA_O	0.2 (0.15)	0.17 (0.03/0.97)
		SA_1	0.18 (0.15)	0.15 (0.03/0.95)
		SA_2	0.22 (0.15)	0.18 (0.07/0.97)
		SA_3	0.21 (0.16)	0.17 (0.07/0.97)
	M_3	SA_O	0.17 (0.13)	0.13 (0.03/0.97)
		SA_1	0.15 (0.09)	0.13 (0.03/0.78)
		SA_2	0.17 (0.11)	0.13 (0.03/0.58)
		SA_3	0.21 (0.19)	0.15 (0.05/0.97)

Table G.8: Within model comparison statistics (DOF = 1) of SA probe response time (minutes) by SA level.

	Decision Difficulty	Sample Size	Mann-Whitney-Wilcoxin
IA	SA_O	670	U = 64732, $\rho < 0.001$
	SA_1	281	U = 10750, $\rho = 0.2$
	SA_2	224	U = 7888.5, $\rho < 0.001$
	SA_3	165	U = 3721.5, $\rho = 0.3$

Table G.9: Spearman correlation analysis between SA probe response time (minutes) and SA probe accuracy by SA level.

	Decision Difficulty	IA Correlation
M_2	SA_O	r = -0.16, $\rho < 0.01$
	SA_1	r = -0.18, $\rho = 0.03$
	SA_2	r = -0.34, $\rho < 0.001$
	SA_3	r = 0.03, $\rho = 0.78$
M_3	SA_O	r = -0.21, $\rho < 0.001$
	SA_1	r = -0.24, $\rho < 0.01$
	SA_2	r = -0.2, $\rho = 0.03$
	SA_3	r = -0.12, $\rho = 0.28$