Human-Swarm-Teaming Transparency and Trust Architecture

Adam J. Hepworth, Daniel P. Baxter, Aya Hussein, Kate J. Yaxley, Essam Debie,
Hussein Abbass *Fellow, IEEE*
School of Engineering and Information Technology
University of New South Wales
Canberra, Australia

***Abstract*—Transparency is a widely used but poorly defined term within the explainable artificial intelligence literature. This is due, in part, to the lack of an agreed definition and the overlap between the connected — sometimes used synonymously — concepts of interpretability and explainability. We assert that transparency is the overarching concept, with the tenets of interpretability, explainability, and predictability subordinate. We draw on a portfolio of definitions for each of these distinct concepts to propose a Human-Swarm-Teaming Transparency and Trust Architecture (HST3-Architecture). The architecture reinforces transparency as a key contributor towards situation awareness, and consequently as an enabler for effective trustworthy Human-Swarm Teaming.**

***Index Terms*—Artificial Intelligence, Explainability, Human-Swarm Teaming, Interpretability, Predictability, Swarm Shepherding, Transparency**

I. Introduction

Recent technological advances in Artificial Intelligence (AI) have made the realisation of highly autonomous artificial agents possible. Contemporary robotic systems demonstrate their increased capability in tasks that were previously exclusive for humans, such as planning and decision making. The advent of swarm robotics furthered the potential of robot systems through the utilisation of a group of relatively simple robots to achieve complex tasks that cannot be achieved by a single, sophisticated robot [1]–[3]. For example, the distributed nature of a swarm of robots gives the swarm an ability to be in different locations at the same time, something a single robot can't do. This could be useful in moving a large object or simultaneous sensing of a large area.

Taking inspiration from biological swarms, robot swarms use local sensing and/or communication and simple agent-logic to achieve global, swarm-level behaviours [4], [5]. A swarm offers more advantages including physically and computationally smaller robots, robustness against failure, and flexibility. In general, robot swarms can be categorised by three broad properties of being flexible, robust, and scalable [6].

Swarm systems are still lacking the human-like intelligence abilities required to manage novel contexts [7]. The performance of fully autonomous swarms is more sensitive to environmental conditions than human-swarm teams [8]. For the foreseeable future, involving the human element into swarm operations is deemed necessary [9]. Nonetheless, the integration of such highly autonomous entities brings new requirements beyond those present in classic master-slave design-philosophy where a machine was to execute only commands issued by its human operator [10].

One of the main requirements that enables task delegation in such team settings is trust [11]. Trust was shown to be an influential variable with a causal effect on human reliance on swarm [12]. Previous findings suggest that when trust is based only on swarm capability, humans run the risk of over-reliance on swarm [13]. Meanwhile, when human trust is also based on an understanding of swarm operation, this trust enables proper task delegation without dismissing human ability to intervene with swarm operation in case of errors [13]. These experimental results demonstrate *transparency* as the necessary base ingredient for trust, with *reliability* providing the ability to improve trust over time, which is consistent with well-recognised models for human trust in automation (e.g. [14], [15]). Trust will likely be vital for ensuring effective collaboration in human-swarm teaming (HST) systems.

This paper proposes a trust-enabled transparency architecture for HST, we call: Human-Swarm-Teaming Transparency and Trust Architecture (HST3-Architecture). The architecture is based on the hypothesis that maintaining a high level of situational awareness (SA) is an enabler for human decision making [16] and a facilitator of appropriate trust [13], which in turn is essential for effective HST. As such, we decompose transparency into three tenets, which, when applied, support the human in improving their SA of swarm actions, behaviours and state information. Transparency in the human-machine systems literature has commonly been situated in collaborative team settings, hence its direct relation to trust and improved joint performance [17]. We present the architecture in the context of HST, where the cooperative attributes of the interaction are highlighted. Nonetheless, the architecture can be equally employed in other HST settings (e.g. [9]) and under different degrees of cooperation where the resulting transparency might or might not be utilised for shared human-swarm goals.

The HST literature is still in its infancy. The tenets for transparency have been discussed in other fields that we will refer to as HxI, such as Human-Swarm Interaction, Human-Robot Interaction, Human-Autonomy Interaction, and Human-Computer Interaction. We will more often in this literature review draw on the literature in HxI to put forward the requirements for effective HST.

We begin by conducting a review and critique of the current literature, covering HST and the fundamental tenets for transparency of interpretability, explainability, and predictability, contained within Section II. Following this, Section III introduces our proposed architecture to realise transparency for HST, intending to promote higher SA. We demonstrate the use of this architecture for a specific case of swarm control, shepherding a flock of sheep through the use of a drone, known as Sky Shepherding. We then present critical areas of open research for the proposed architecture in Section V and conclude the paper in Section VI.

II. Literature Review

A. Transparency

AI systems are generally categorised into two types: black-box and white-box. Black-box refers to a system in which the inputs and outputs can be easily identified, but how the outputs are derived from the inputs is unknown. White-box (also known as glass-box) refers to a system whose internal algorithmic components and/or its generated model can be directly inspected to understand the system's outputs and/or how it reaches those outputs [18]. Examples of such systems include decision trees [19], rule-based systems [20], [21], and sparse linear models [22]. The white-box category is generally accepted as more transparent than the black-box one.

Transparency is an essential element for HxI, yet is also a concept with significant variations in definition, purpose and application [23]. For example, in ethics, transparency is the visibility of behaviours, while in computer science, it often refers to the visibility of information [24]. For agent-agent interactions, transparency is used to assist with decision making [25]. For an autonomous agent, be it biological or artificial, transparency has attracted the interest of many researchers due to its facilitation role in team collaboration [17], [25], [26].

However, there is little research focusing on swarm transparency, due in part to the recent emergence of HST as a distinctive area. Additionally, the unique challenges of swarm systems may have impaired further research advancement for swarm transparency [27]. One of the main challenges is the decision of whether transparency is needed on a micro or a macro level. Micro-level transparency exposes information about the state of each swarm member, which can be useful in identifying failed or erroneous entities [28]. However, for large-sized swarms, micro-level transparency could impose significant bandwidth requirements beyond what is reasonably possible [29]. Also, the amount and level of information can overwhelm a human, limiting their capability to keep track of what is going on [30]. Macro-level transparency is useful for offering an aggregate picture for the state of the swarm, but comes at a cost in opacity, obscuring many low-level details. Within the literature, experimental results are divergent for which transparency level is *better*, even for basic swarm behaviours [30].

A further challenge for swarm transparency stems from the fact that global swarm behaviours emerge from the actions of its individual members, with knowledge and behaviours of swarm individuals being locally focused. Typically, swarm members are assumed to be unaware of the global state of the swarm [4], and hence, of whether their behaviours align with the desired collective swarm state. Consequently, swarm members might not be able to provide satisfactory explanations for their actions, the collective behaviour of the swarm, or importantly understand their role or task within the collective.

Another challenge for transparency is to consider how to support everyday interactions between human-machine. In [11], the authors proposed using cyber to support such interactions, leading to the possibility of swarms existing beyond the physical. The adaptability, robustness, and scalability of swarm systems are also inspiring research into abstract modelling of cyber-physical systems to support understanding complex problems [31], [32]. Swarms and swarm behaviour can exist in both the physical and cyber realm [27] and will require varying levels of human interaction.

The physical state of swarm individuals such as position, battery level, and damage, can be aggregated to give a simplified view of a swarm member's physical state. What remains less clear is how the virtual state of swarm members, for example, confidence levels [33] or intra-swarm trust [34], can be communicated without overloading the human. In collective decision-making problems, calculating a mere average of the confidence levels does not provide an answer to which members influence the decision making process or whether the swarm is expected to converge on a correct final decision.

Transparency has received a great deal of researchers' interest across various fields. The quest for transparency entails the answer to two questions: 1) What are the desirable aspects of transparency? And, 2) How to achieve these aspects? Endsley's SA model [35] defines what levels of knowledge a human should maintain to enable successful interaction with their automation teammate. Chen's model for agent transparency (SAT) maps these levels into corresponding aspects of transparency that are required to be exhibited by the automation. Each level consists of similar goals to [35] to enable transparently shared understanding [36] by articulating *what* information should be conveyed at each level.

When using the SAT model to assess trust factors of transparency and reliability, Wright et al. [37] found that SAT was able to support human decision making regardless of the reliability of the autonomous agent. However, the human was unable to reconcile trust after observing erratic behaviour by the autonomous agent, regardless of SAT level used during task completion. Consequently, while the SAT model supports transparent decision making, it is unable to support trust relationships in moments of unreliability due to its emphasis on the "what" rather than the "how" to architect a transparent system.

The SAT model is thus helpful in defining what sort of information is necessary for each transparency aspect and when each aspect should be made available. Interpretability is key for understanding, explainability is needed for comprehension, while predictability is required for projection. These are the three tenets of transparency required to support SA and SAT. However, the SAT model does not specify how to engineer these tenets which is the gap our architecture aims to address.

Unfortunately, there is inconsistency in using these concepts in the literature. The remainder of this literature review section is structured around each of these three concepts. We aim to reduce the inconsistency around the multiple, often used synonymously, confusing definitions for transparency. The literature survey considered work in technological fields, including robotics, computer science, and swarm research. When further grounding of terms is necessary to reduce ambiguity, we draw on psychology, linguistics, and human factors. We organise the remainder of this section according to the three tenets of transparency, being interpretability, explainability, and predictability.

B. *Interpretability*

Interpretability in artificial intelligence is a broad and poorly defined term. Moreover, the present state of the interpretability literature in the context of swarms is limited. Generally speaking, to interpret means to extract information of some type [38].

The literature differentiates two types of interpretability, being algorithmic and model. Algorithmic interpretability is the ability to inspect the structure and hyper-parameters of a system to understand how it works. This is useful to answer questions about the algorithmic component of AI systems: i.e. does the algorithm converge? Does it provide a unique decision? Is the role of its hyper-parameters well-understood? Model interpretability is related more to the model learned by the algorithmic component and used to map inputs to outputs. Several non-mathematical definitions exist in the literature for model interpretability, such as Miller who states that “ interpretability is the degree to which an observer can understand the cause of a decision” [39, pg.8]. Kim et al. who states that “a method is interpretable if a user can correctly and efficiently predict the method’s results” [40, pg.7], and Biran and Cotton who state “systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation” [41, pg.1]. These definitions convolute the concept of interpretability, causality, explainability, reasoning, predictability and anticipation.

Only a few studies in the literature have investigated interpretability concepts in the context of swarming systems. In this regard, interpretability is used to express communication among agents. For example, Sierhuis and Shum [42] developed a conversational modelling tool that is used in realistic analogue simulations of collaboration between humans on Earth and robots on Mars, referred to as Mars-Earth scientific collaboration. Lazaridou et al. [43] investigated interpretability in scenarios where agents learn to interact with each other about images. Das et al. [44] examined interpretability in his study where agents interact and communicate in natural language dialogue on a cooperative image guessing game. The agents recognise the contents of images and communicate that understanding to the other agents in a natural language. These communicating agents can invent their communication protocol and start using specific symbols to ask and/or answer certain patterns in an image. The agents then leverage a human-supervised task to structure the learned interaction in an understandable way for human supervisors. Andreas et al. [45] also examine messages exchanged between agents using learned communication policies. A strategy is developed to translate these messages into natural language based on the underlying facts inferred from the messages. St-Onge et al. [46] studies the expressiveness of swarm motion as a way to convey high-level information to a human operator. The swarm motion is tuned to share different types of information. Swarm aggregation, graph formation, cyclic pursuit, and flocking are examples of motions used to express different information. Suresh and Martínez [47] developed an interpreter, an interface between the human and the swarm, that takes in high-level input from a human operator in the form of drawn shapes and translates it into low-level swarm control commands using shape morphing dynamics (SMD). Further, the interpreter is also used for translating feedback to a human operator.

The work on interpretability characterises the behaviour of AI systems in terms of their architecture (or algorithmic components), learned computational models, goals, and actions. Despite the extensive work done in this regard, these definitions do not explicitly account for the capabilities and characteristics of the observer agent and its capacity to recognise and synthesise the interpretations provided. An agent’s behaviour may be uninterpretable when it does not comply with the assumptions or the cognitive capabilities (i.e. knowledge representation, computational model, or expertise) of the observing agent [48].

C. *Explainability*

Explainability is required for trust, interaction, and transparency [49], although knowing precisely what is needed for a good explanation remains unclear [50]. As Minsky notes, humans find it hard to explain meaning in things because meaning itself depends on the environment and context, which is distinct for every person [51]. We reason on human-understandable features of the inputs (data), which is a critical step developing the chain of logic of how or why something has happened or a decision was made [52]. Humans can learn through a variety of methods and transfer experiences and understanding from one situation to another develop heuristic short-cuts, like common sense, along the way. Complex chains of reasoning with ill-defined elements tend to make it difficult to explain and justify decisions [49]. An explanation can be developed dynamically after the fact, becoming communicated through a story from a mental model developed in the mind of the person communicating the explanation [53].

Many definitions for explainability have been published, overwhelmingly without the precision of a mathematical accompaniment. Explainability definitions vary substantially in terms of length, ambiguity, and context. Many are relatively short yet insightful, such as Josephson and Josephson who state that “an explanation is an assignment of causal responsibility” [54, pg.14] or that “an explanation is the answer to a why-question” [55, pg.7]. The Defense Advanced Research Agency (DARPA) explainable artificial intelligence (XAI) program indicates that XAI “seeks to enable third-wave AI systems, developing machines with an enhanced understanding of the context and environment for which they operate in” [56, pg.1], although without bounding the problem space further.

Miller states that “explanation is thus one mode in which an observer may obtain understanding, but clearly, there are additional modes that one can adopt, such as making decisions that are inherently easier to understand or via introspection” [55, pg.8] or from the perspective of cognitive architectures “information is a linguistic description of structures observable in a given data set” [57, pg.3].

The previous definitions are generally considered too broad or ill-defined to enable adoption of a formal architecture. Application of such definitions as those selected here requires an implicit input from the user, particularly their expertise, preferences, and environmental context [58].

Nyamsuren and Taatgen [59] argue that human general reasoning skill is inherently ‘a posteriori’ inductive, or probabilistic. They base this on two key points: 1) deductive reasoning, in its classical form, states that what is not known to be true — is false, which therefore assumes a closed world; and, 2) humans have shown to use an inductive, probabilistic reasoning process even when seemly reasoning with deductive arguments. This world view can be on a local- or global-level, described as either micro or macro from the systems perspective. Local explainability refers to the ability to understand and reason about an individual element of the system, such as a particular input, output, hyper-parameter, or algorithmic component if the system consists of more than one. Several definitions exist for explainability that can be categorised as a local explanation, such as Miller [39] who state that local explanations detail a particular decision of a model to determine why the model makes that decision. This is commonly achieved by revealing casual relations between the inputs and outputs to the model. Biran and Cotton state that a justification “explains why a decision is a good one, but it may or may not do so by explaining exactly how it was made. Unlike introspective explanations, justifications can be produced for non-interpretable systems.” [41, pg.1] Global explainability often describes an overall understanding of how a system functions or an understanding of the entire modeled relationship between inputs and outputs [60]. A system is said to be globally explainable if its entire decision-making process can be simulated and reasoned about by an external agent, who is a target for the explanation [61].

Long standing questions around how to produce an explanation, if a process should be explainable [62] and what should be required for an explanation to be considered sufficient [50] remain open within the literature. These notions follow from what Searle described as within the realm of strong-AI [63], noting that machines must simulate not only the abilities of a human but also replicate the human ability to understand a story and answer questions. A desire for machines to imitate and learn like humans is not a new concept [64].

While reasoning presents itself as a method for a logical explanation, fundamental questions of *What, Who, Which, When, Where, Why* and *How*, i.e. the ‘wh’ -clauses, of explainability require careful consideration. Rosenfeld and Richardson [65] highlight the interconnecting nature of these questions and assert the motivation for the system itself has a direct bearing on the overall reason or reasons the system must be explainable. Whether the system is designed as human-centric, built to persuade the human to choose a specific intention, action, or outcome; or, agent-centric, to convince the human of the correctness of their intention, action or outcome, the explanation provided should contribute to the overall transparency of the system—including the human. Explaining is far more effective when a co-adaptive process is employed [66], which Lyons [67] discusses through an HxI lens as robot-to-human and robot-of-human factors. Only then can one determine *What* explanations are required, *Who* the explanations are directed toward, *Which* explanation method suits, *When* the information should be presented or inducted, *Where* they should be presented or inducted, *Why* explainability is needed in the system, and *How* objective and subjective measures can be used to evaluate the system [65].

There are many swarm system control mechanisms and architectures that have been developed and introduced, but insufficiently address understanding for supervisory control of such systems, particularly for the human interaction with various levels of swarm autonomy [68]. Previous research has investigated the principles of swarm control that enable a human to exert influence and direct large swarms of robots. What has been lacking is the inclusion of bi-directional, interpretable communication between the supervisor and the swarm. This has limited the development of a shared understanding as to why or how either the human or swarm is making decisions. Addressing such information asymmetry is essential to realise HST [69] fully. Such asymmetry manifests during HST where some actions or behaviours may not be immediately apparent to the human if a swarm behaviour doesn’t align to the human’s expectations [69]. The swarm may assess that this behaviour is optimal to achieve the goal, but requires explaining to the human in order to ensure that confidence in the swarm is maintained. Previous HST studies have noted the importance of appropriate and consistent swarm behaviours, although lack a method to provide feedback to the human [70]. This asymmetry of information highlights the difference between human-to-swarm and swarm-to-human communication, which is an essential element to consider for facilitating teaming. Swarms are commonly used to support a human’s actions in HST [70]. However, without an ability to query how the swarm chooses a future state, human team members may not be confident in action being taken by the swarm. In such situations, interrogation of the swarm to generate explanations could alleviate such issues of confidence, increase shared understanding, and build trust for HST.

D. Predictability

The Cambridge Dictionary defines the word predict as “to say what you think will happen in the future” [71]. The term “predict” has been used by researchers to refer to not only forecasting future events but also estimating unknown variables in cross-sectional data [72]. Similarly, the term “predictability” is used to denote different notions including: the ease of making predictions [73], behavior consistency [74], and the variance in estimation errors [75].

An agent’s predictability has received significant attention in HxI due to its significant impact on interaction and system-level performance. Agent predictability has been defined as the degree to which an agent's future behaviours can be anticipated [14]. The term has also been commonly used to reflect the consistency of an agent's behaviour over time (e.g. [76]–[78]). Coupling predictability and consistency implicitly assumes that making future predictions about an agent is completely performed by the agent's teammate (or observer) and that these predictions are heavily based on historical data and/or pre-assumed knowledge.

The ability to predict an agent's future actions and states is a crucial feature that facilitates the collaboration between agents in a team setting [79]. In highly interdependent activities, predictability becomes a key enabler for successful plans [79]. Also, predictability is an essential factor that facilitates trust by ensuring the matching between the expected and received outcomes [80]. Previous research utilised predictability to achieve effective collaboration in various HxI applications including industry [81], space exploration [82], and rehabilitation [83]. Depending on the requirements and the issues present in these domains, agent predictability was aimed to serve different purposes that can be grouped into the following areas: mitigating the effects of communication delays, allowing humans to explore possible courses of actions, enabling synchronous operations and coordination, and planning for proactive collision avoidance.

Remote interaction between agents can be severely impacted by considerable communication delays that impede their collaboration. This is particularly the case for space operations where the round trip communication delay is several seconds [84]. Such a delay was shown to be detrimental as it negatively affects mission efficiency and the stability of control loop [85]. One of the earliest and most widely used solutions to mitigate the effects of significant communication delays is the use of predictive displays. A predictive display uses a model of the remote agent, its operation environment, and its response to input commands to estimate the state of the mission based on the historical data recently received from the agent. This enables predictive displays to provide an estimation of the current state of the agent that has not been received and to provide timely feedback on the predicted future agent's response to input commands that is yet to be received by the agent. This allows for smooth teleoperation as compared to the inefficient wait-and-see strategy [86]. Past studies show that predictive displays can maintain mission performance [87] and completion time [88] at levels similar to no delay. Previous findings also demonstrate the effectiveness of predictive displays in enhancing the concurrency between remotely interacting agents under variable time delays [89].

Robot predictability is the main subject of investigation in studies involving predictive displays that are either used to account for communication delays [88] or to facilitate exploring action consequences [81]. Likewise, robot predictability is the focus when people are assumed the responsibility for avoiding collision with the robot [90]. As for systems where a human executes a physical activity that needs to be synchronised with robot actions, human predictability becomes an enabler for successful operation. This can be the case for some industrial applications [91] or rehabilitation scenarios [83], [92]. While predictive displays are mainly used to enable effective interaction in the presence of considerable time delays, the same concept has been used to allow for the exploration of the consequences of user commands without actually executing them. Several studies proposed the use of predictive displays to predict robot responses to human commands for training [81], validation [84], and planning [93] purposes. In such cases, human input commands are sent only to the virtual (simulated) robot and not to the actual robot. This enables people to explore how their actions affect the state of the remote robot without causing its state to change. Once the human is satisfied with the predicted consequences of a command (or a sequence of commands) and the command passes the essential safety checks, it can be committed and sent to the remote robot to execute.

Another crucial purpose for agent predictability is to enable the synchronisation and coordination between collaborating agents. Action synchronisation can be critical for the success of highly interdependent tasks. For instance, there are studies which investigate the utility of using predictions about humans' intended future actions to enable the operation of assistive and rehabilitation robots [83]. These predictions can then be used to calculate the optimal forces a robotic limb should apply to help the human perform the intended movement without over-relying on the robot [92].

Collision avoidance is also an area that benefits significantly from agent predictability. While operating within an environment shared with other moving objects, an agent needs to ensure collision-free navigation to avoid possible damages or safety accidents. A fundamental way proposed to use agent predictability for collision avoidance was to require the agent to announce its planned trajectory so that other agents can plan their motion accordingly to avoid it [90]. Other approaches focus on equipping the agent with the ability to detect the motion and predict the future positions of other moving objects so that the agent can actively act to avoid a collision. The agent may not be able to plan a complete collision-free path from the onset. Instead, the agent can continuously monitor its vicinity and predict whether the motion of the other agents will intersect with its planned path causing possible collisions [94], [95]. This allows the agent to proactively avoid collision by re-planning its path according to its updated prediction, or wait till the path is clear if necessary.

Swarm predictability has been the focus of only three papers [87], [96], [97]; all of which report on the same experimental study. In that study, the predicted state of swarm members is used to enable human control of the swarm under significant time delays. Besides predicting agents performing the task, task success can also necessitate predicting the state of other agents or objects that share the same operating environment. For instance, human bystanders are the agents to be predicted in systems where the robot has to ensure collision-free navigation in its path planning [94], [95].

III. Human-Swarm-Teaming Transparency and Trust Architecture (HST3-Architecture)

A. Design Philosophy

Autonomous systems will continue to increase their level of smartness and complexity. These desirable features, necessary

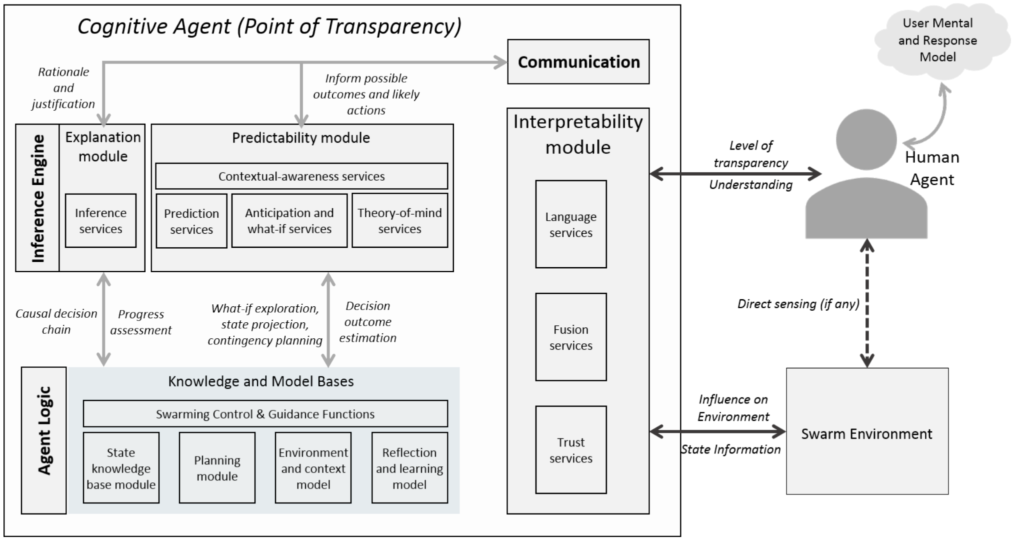
Cognitive Agent (Point of Transparency)
Rationale and justification
Inform possible outcomes and likely actions
Communication
User Mental and Response Model
Inference Engine
Explanation module
Inference services
Predictability module
Contextual-awareness services
Prediction services
Anticipation and what-if services
Theory-of-mind services
Interpretability module
Language services
Fusion services
Trust services
Level of transparency
Understanding
Human Agent
Causal decision chain
Progress assessment
What-if exploration, state projection, contingency planning
Decision outcome estimation
Direct sensing (if any)
Agent Logic
Knowledge and Model Bases
Swarming Control & Guidance Functions
State knowledge base module
Planning module
Environment and context model
Reflection and learning model
Influence on Environment
State Information
Swarm Environment

Figure 1. The generalised HST3-Architecture.

for autonomy in complex environments, bring the undesirable effects of making their teaming with humans significantly more complex. As we expand the teaming arrangements from a single autonomous system to a swarm of autonomous systems, the cognitive load of the humans involved increases significantly; thus leading to unsuccessful teaming arrangements. Two design principles are needed to reduce the cognitive overload from their teaming with a swarm. First, transparency in swarm operations is required to enhance contextual awareness by understanding changes in priorities and performance. Second, two-way interaction between a human and swarm offers the ability for a human operator to ask questions and receive answers during swarm operations at different points in time to revise goals, beliefs, and update operating conditions.

HST3-Architecture promotes swarm transparency, while integrating two-way interaction in HST systems. HST3-Architecture leverages the proposed tenets of transparency and Endsley's SA model [35] to develop a generic transparency system to maintain a mutual and/or shared understanding of the current status, plans, performance history, and intentions between human operators and the swarm.

In contrast to the focus on the *what* in the SAT model [25], we propose an architecture that is focused on *how* by adopting a systems engineering approach to fuse together interpretability, explainability, and predictability. By linking the tenets to transparency together, we can design for both transparency assurance, as well as diagnosis for failure tracing in a system. We first focus on a functional definition of the three tenets of transparency: interpretability, explainability, and predictability.

Interpretability allows for shared knowledge of the situation and outcomes [55], [98]. Interpretability supports transparency by ensuring that knowledge is transferred properly among agents. Interpreting [99] is the process of mapping spoken words between two languages. Consequently, interpretability enables transparency by facilitating communication between agents using a knowledge set that includes language and processes. Interpretability could be seen as a form of translation to convey original meaning [99], could include sentiments [100], and capture cognitive behaviours such as emotions [101].

While some authors use the terms explainability and interpretability interchangeably [55], we contend that the two terms must be differentiated. By positioning interpretability as a functional layer between the system's ability to explain and the agents a system is interfacing with, we eliminate ambiguities and achieve a modular design for autonomous systems that separate the two functions. Explainability augments interpretability with deeper insights into sentiments and an agent's cognitive and behavioural states by expanding the causal chain that led to the state that is subjected to interpretability. Explainability assigns understanding to an observer's knowledge base by providing the causal chain that enables the observer to comprehend the environment and context it is embedded within. Comprehension of the environment allows the observer to improve their SA and support robust decision making [16].

Interpretability and explainability together offer an observer with understanding and comprehension of a situation. The sequence of transmission of meaning to an observer affords the observer with necessary updates in the observer's knowledge base. These knowledge updates are necessary for the observer to infer whether or not the sequence of decisions is expected. The updates allow agents to deduce consistency of rationale and induce or anticipate future actions. Such consistencies promote mutual understanding of an outcome [102].

The knowledge updates achieved through interpretability and explainability form the basis for predictability. As a necessary component for joint activity [79] and team success, mutual predictability becomes an engineering design decision facilitated through explicitly defined procedures and expectations. Predictability among agents brings reliability [80] to transparency. By using transparency as the basis of our architecture, and enabling reliability by design, we present an architecture that supports human-swarm teaming and offers a modular design to inform trust calibration.

Figure 1 presents a conceptual diagram of the architecture.

The direct line of communication to the swarm is through the user-interface, and therefore becomes the focal point of the outputs produced by the interpretability, explainability and predictability modules. For efficient HST, the system should only exchange with external actors through the interpretability module. System explanation and prediction information are parsed to the interface once mapped into a human interpretable format by the interpretability module. Additionally, the user module can query the system, through the interface, at any time for state information, explanations, and/or prediction requests.

B. Disambiguating the Tenets of Transparency

The literature review has demonstrated the confusion in the existing literature on appropriate definitions for the tenets of transparency. Before we are able to present an architecture for transparency that encompasses these tenets, it is pertinent that we disambiguate these concepts by presenting concise definitions.

- A_w: A worker agent whose logic is required to be interpreted to another agent.
- A_o: An observer agent that synthesises the behaviour of a working agent (A_w) to understand its logic.
- E: The environment that provides a common operating context for both agents A_w and A_o.
- $S(t)$: The state of the mission (i.e. overall aim the human and the swarm aim to achieve together) at time t.
- $S_w(t)$: The state of the mission as perceived by the worker agent at time t.
- $S_o(t)$: The state of the mission as perceived by the observer agent at time t.
- L_w: Algorithmic components of the working agent.
- M_w: The computational model learned by the worker agent through interactions within E.
- M_o: The computational model of the observing agent.
- K: A knowledge set.
- α and γ: internal decisions made by an agent.
- $\prec$: A partial order operator.
- $\mathbf{\Omega}$: A decision that has been made internally and expressed externally by an agent.

Definition 1. Interpretability, I, is a mapping of a system's behaviour in terms of its algorithmic components, computational model and mission state, to a knowledge set K in a form appropriate for observing agent to integrate with its internal knowledge (i.e. context, goals, intentions, and computational capabilities of the observing agent).

$$I\colon\ (L_w, M_w, S_w(t)) \rightarrow K \tag{1}$$

Definition 2. Explainability, $\mathbb{E}$, defined as a sequence of expressions in one language that coherently connects the inputs to the outputs, the causes to the effects, or the sensorial inputs to an agents' actions.

$$\mathbb{E} : (L_w, M_w, S_w(t), \mathbf{\Omega}) \rightarrow \alpha \prec \gamma \prec \cdots \rightarrow \mathbf{\Omega}, \tag{2}$$

Definition 3. Predictability, P, is an estimation of the next state of a mission given previously observed mission states by an agent.

$$P : (L_w, M_w, S_w(t : t - \tau)) \rightarrow S(t + 1), \tag{3}$$

C. The Architecture

Figure 2 illustrates the architecture for the proposed transparency and trust architecture. HST3-Architecture follows a three-tier architecture and is typically composed of an agent knowledge tier (lower layer), an inference engine tier (middle layer), and a communication tier (top layer).

The lower layer of HST3-Architecture contains state information on task-specific knowledge and the learning processes used by the agent. The middle layer consists of two primary modules being explanation and prediction. The explanation module presents to the operator's the causal chain of events and state-changes that led to the current state of both individual swarm members and the swarm as a whole. The predictability module supports projection and anticipation functions by informing the swarm's future states and what-if analysis. The top layer is a bidirectional communication layer that interprets messages exchanged between the human operator and the swarm. It interprets the swarm's state information, reasoning process, and predictions in a langauge and framing calibrated to the human operator. It also maps a human operator's requests into appropriate representations commensurate with the swarm internal representations, knowledge, and processes for explanation and predictability.

In the remainder of this section, we will expand on each of these modules.

1) Interpretation Module: The interpretability module acts as the interface with external entities to the swarm and offers bidirectional communication capabilities between the swarm and external human and non-human actors. According to Equation 1, interpretability should account for the computational models for both the working and observing agents, the shared context between both parties, and maintains a knowledge representation method accepted for them. The interpretability module relies on three types of services:

- Language Services: different ontologies, taxonomies, parsing, representations and transformations need to exist to allow the interpretability module to offer bi-direction communication capabilities. In heterogenuous swarms, it could be necessary to communicate in different languages within the swarm, as well as to the actors the swarm is interfacing with.
- Fusion Services: the ability to aggregate and disaggregate information is key for the success of the interpretability module. The swarm need to be able to take a request for a swarm-level state information and decompose it into primitive state information that needs to be fused to deliver the information on the swarm level. These fusion services need to be bi-directional; that is, they are aggregation and de-aggregation operators.
- Trust Services: the ability of the interpretation module to respond to bi-directional communication requests rests

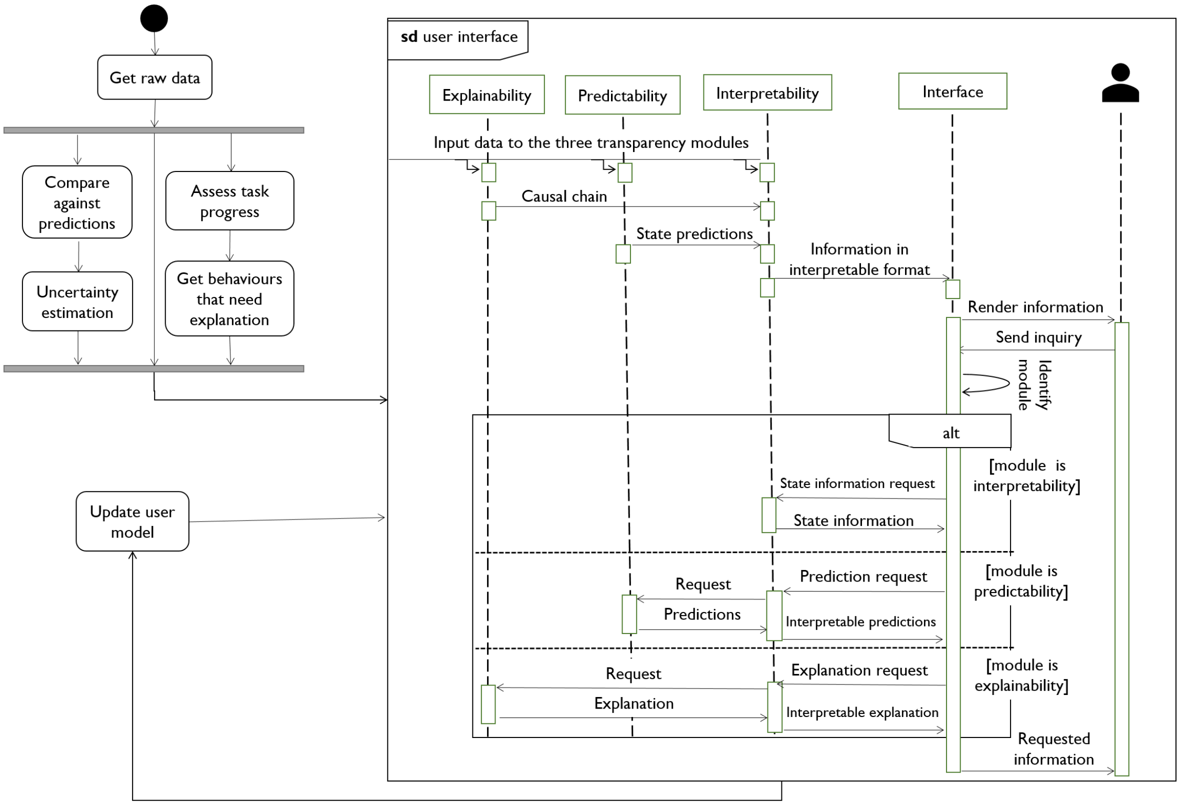

Figure 2. A UML interaction overview diagram describing the HST3-Architecture.

on its ability to represent which information types and contents are permissible by the swarm, allowed for sharing to whom, in which context, and by which member of the swarm. The challenges as well as opportunities in a swarm are that these trust services could de-centralised. All members in the swarm need to have the necessary minimum information needed to allow them to perform trust services during interpretation.

2) Explanation Module: Explainability provides the inference and real-time reasoning engines, as well as the knowledge base within HST3-Architecture. It generates the causal chain output in response to a particular request through the interpretability module as shown in Figure 2. The role of the explainability module is to deliver a service to the user; thus, the actor the swarm is interacting with is a central input to how the swarm should calibrate its ability to explain to be suitable for that particular actor.

While the explainability module responds to a request that arrived through the interpretability module, the output of the explainability module needs to go back through the interpretability module to be communicated to the actor(s) interacting with the swarm. The unification of the representation and inferencing mechanisms within the explainability module offers an efficient mode of operations for transparency. For example, a single neural network to operate, metaphorically a single brain, while allowing the swarm, through the interpretability module, to, metaphorically, speak in different languages.

The explainability module offers the user with understanding of the behavioural decisions, past, present, and future, of the swarm through the single-point of explainability. The user receives an explanation of the current state of the swarm, answering questions such as: what are the swarm members doing? Why are they doing it? What will they do next? That is, the type and level (micro vs macro) of the explanation information will be driven extrinsically by user inquiries and intrinsically by trust and teaming calibration requirements.

3) Predictability Module: Predictability is a bi-directional concept. First, the swarm needs to be able to share their information to eliminate surprises for the human. Second, the swarm needs to be able to anticipate human states and requirements. The simplest requirement for predictability is to respond to questions requiring projection of past and current states into a future state. Such a requirement can be achieved in its basic form through classic prediction techniques.

The performance of the predictability module relies on a few services to deliver mutual predictability in the operating environment; these are: contextual-awareness services, prediction services, anticipation and what-if services, and theory-of-mind services.

Mutual predictability requires the swarm to maintain situation awareness on the context; that is, the state of itself, other actors, and their relationship to the overall objectives that need to be achieved in this environment. The decentralised nature of the swarm means that the context is distributed and needs to be aggregated from the information arriving through the interpretability and explainability modules. The contextual-awareness services depend on prediction services, anticipation and what-if services, and theory-of-mind services.

The anticipation and what-if services simulate in the background the evolution of the swarm and human-swarm interaction to anticipate physical and cognitive state variables. These simulations run in a fast-time mode, projecting future evolutionary trajectories of the system to identify what is plausible. The emphasis of anticipation is more on plausibility and less on prediction. Meanwhile, the prediction services focuses more on prediction. These services are more data-driven than model-driven and simulation-driven. They rely on past and current state-information to estimate future states.

Mutual predictability requires an agent to have a model of itself and other agents in the environment. These models are offered in our architecture using the theory-of-mind services, which model other agents in the environment with an aim to anticipate their future needs of information to ensure these pieces of information are communicated to improve mutual predictability.

Human‘s form their internal mental models based on observations and previous knowledge of different situations [103]. A mental model, as defined in the cognitive psychology literature, is a representation of how a user understands a system. In the context of swarm transparency, state information, explanations, and predictability on the micro and macro level help to form and update more accurate and complete mental models of the system [104], and the complex algorithmic decision-making processes embedded within the actors [105].

The formation of mental models comes at a cost—an increase in cognitive load. The cognitive load required to build a mental model is dependent on the type of mental model, the complexity, amount, and level of information presented to the user for processing [104]. The four distinct functional abilities Langley et al. [106] further developed from Swartout and Moore’s description [50], can be directly applied to HST through Lyons‘s four transparency models [67]. At the highest level, building an intentional mental model allows the human to understand the intent or purpose of the swarm. Once this is understood, the user can begin to build swarm task mental models. To achieve this, the system must explain which actions the swarm executed and why, the plans and goals the swarm adapted, or inferences the swarm made to the user.

The predictability module has two modes of operation, being an autonomic-mode and an on-request mode. The autonomic-mode perform self-assessment of self-needs and the needs of other actors in the environment, then acts accordingly. The default mode can also be set to send predicted risks that can disrupt swarm operations. The on-request user inquiry mode seeks to provide dynamic predictions, for instance, to answer a question arriving from external entities to the predictability module and user-questions such as “where will the swarm be in five minutes?”, or “what is the predicted battery level of a particular swarm member at some specified time in the future?”. That is, the type and level (micro vs macro) of the predicted information will be driven by external and user inquires.

The adaptability needs to adhere to human cognitive constraints by presenting only the main predicted state variables while also communicating granular predictions as per user needs. The theory-of-mind services play a crucial role in the assurance of this requirement.

D. *Objectifying Transparency*

It is less productive to discuss transparency in a technology purely from a qualitative perspective, without offering designers and practitioners appropriate concrete guidelines and metrics to guide and diagnose their designs. The core motivation for transparency in HST is to improve the efficiency and effectiveness of the overall system-of-systems composed of the swarm and all other actors, including humans, involved in the delivery of the overall solution.

While providing a measure of transparency is essential for the user, what also must be considered is the accountability of this answer. Determining the contribution each of interpretability, explainability, and predictability to situations where the provided information does not satisfy the needs of human operations is an important consideration. Moreover, the measurement and reporting of transparency must consider the level of granularity required against human operators cognitive capacity to ingest, process, and use the information.

The HST3-Architecture offers an advantage through design, by considering the level and type of information presented to the user, per Figure 1. The measurement and evaluation of the HST3-Architecture is an essential element, which enhances transparency by decreasing or eliminating all together any level of opaqueness. Transparency leads to an increased user SA, and ultimately system reliability [107].

The measurement and evaluation of transparent systems have been previously identified as a research gap to understand the user-based measures [58], and how both objective and subjective measures can be used to evaluate a system designed to be transparent [65]. We assert that to measure and evaluate a level of transparency, indicators for each of the transparency tenets must be measurable. Each tenet of transparency in HST3-Architecture could then be evaluated using a multitude of metrics in the literature. For example, several existing objective and subjective measures and meta-categories have been proposed in the literature [58], [107]–[109]. Interpretability could be evaluated using a questionnaire, by asking the user whether the message arriving from the swarm is easy to understand or not. We will use β to indicate a function that outputs one of three levels for each tenet in achieving its intent, where

- β_0, indicating that one of the tenets is either absent or non-functional. For example, $\beta_0(I)$ indicates that the system does not have an interpretability module that is functioning properly.
- β_1, indicating that one of the tenets is functioning at a level deems to be fit-for-purpose for the task. We do not assume, or aim for, perfection due to the fact that every technology is evolving in its performance as the context and environment continues to evolve. For example, $\beta_1(I)$ indicates that the system has an interpretability module that has been assessed to be functioning properly and is communicating in a language appropriate for other agents to understand.

Table I
Cases of Transparency

Case	Transparency Case	I	E	P
1	Opaque	0	#	#
2	Confusing	θ	!0	!0
3	Communicative	!0	0	0
4	Rationally Communicative	!0	!0	0
5	Socially Communicative	!0	0	!0
6	Articulate	1	0	0
7	Rationally Articulate	1	!0	0
8	Socially Articulate	1	0	!0
9	Fit-For-Purpose	1	1	1

- β_θ, indicating that one of the tenets is functioning but there is a level of dissatisfaction with its performance. This could be a low, medium or high dissatisfaction. We do not differentiate between the different levels of satisfaction in this paper as they all indicate that a level of intervention is needed to improve this particular tenets.

The tenets of transparency are not additive. As we will explain below, a system that has a functional explainability and predictability modules will be considered a black-box system if the interpretability module is dysfunctional. We will use a wildcard symbol (#) to indicate a don't care match. For example, $\beta_{\#}(E)$, indicates that we do not care about the level of explainability in this system; that is, regardless of whether it is absent all together, partially functional, or a fit-for-purpose, explainability has no impact on transparency in this particular scenario. When a particular state is excluded, we use the exclamation mark as a negation; that is, $\beta_{!0}(I)$ indicates that the interpretability module is either $\beta_\theta(I)$ or $\beta_1(I)$.

We can now define nine distinct cases of transparency using its three tenets: interpretability (I), explainability (E), and predictability (P). The nine cases are listed in Table I

The first case, opaque transparency, is when the interpretability module is absent or not functioning at all. In this case, regardless of whether the swarm possesses internal abilities to reason or has predictability abilities, the swarm is unable to communicate any of these capabilities with external actors. The external actors could observe the swarm's behaviour, and may develop a level of trust if the swarm performs well and they can anticipate its behaviour, but the lack of interpretability makes the swarm unable to communicate to other entities. In other words, the interacting agent is unable to harness any of the tenets that support transparency. Such opaque swarm may be understood post analysis [18]; however, real-time interaction will be problematic.

The second case, confusing transparency, occurs when the interpretability module is making mistakes. The explainability and predictability modules could be functioning perfectly or generating mistakes on their own, confusing the messages the swarm is communicating even further.

The third to fifth cases occur when the interpretation modules is functioning, even partially, and at least one of the explainability or predictability modules are not functioning. In the case when none of them is functioning, the swarm can communicate state information to other actors in the environment, albeit it may break down from time to time if interpretability is evaluated as β_θ. While a level of mutual understanding among the swarm and humans may evolve, it will likely be limited, which will hinder the situation awareness of the agent. An example of this system is presented in [110]. If the explainability or predictability modules function, the case of transparency is called rationally and socially communicative, respectively.

The sixth to eighth cases of transparency mimics the third to fifth cases, except that the interpretability module is fit-for-purpose, thus, it delivers intended meaning consistently. We label this case as an articulate swarm. When either the explainability or predictability module are functioning, the case is labelled rationally and socially articulate, respectively.

The last case of transparency is when all three modules are functioning at a level appropriate for the human-swarm team to operate effectively and efficiently. This a fit-for-purpose transparency.

It might be worth separating an overlapping case that we call Misaligned transparency, when the predictability module is absent, while the interpretability and explainability modules are functional, albeit they may break-down from time to time. In this case, the system is able to communicate its states and causal chains for its decision, but it can't anticipate the states and/or rationale of the actors it is interacting with; thus, the communication will likely get misunderstood sometimes and a level of inefficiency will continue to exist in this system's ability to communicate with other actors in the environment.

IV. A Case Study on Sky Shepherding

We present a case study that describes how our proposed architecture could be applied to a real-world situation. The scenario we use is that of shepherding, which is a method of swarm control and guidance. We consider an environment with three agent types, being a cognitive agent (the human shepherd), the herding agent (a human drone pilot), and the constituent swarm member agents (sheep in a flock).

The Sky Shepherd case is a live example of our current work, where we employ the HST3-Architecture to guide the design of the system, replacing the human pilot and the drone with a smart autonomous drone guiding the swarm and teaming with the farmer in a transparent manner.

In this scenario, our shepherd provides a general direction to the drone pilot. The pilot interprets this direction and begins to plan their tasks, sub-goals, path, and consider the reaction of the flock, employing the agreed knowledge base to understand what future states may look like. After making an assessment and determining the optimal behaviour profile to meet the shepherd's intent, the drone pilot commences their sequence of behaviours towards the sheep and begins shepherding. During the task, the drone pilot agent receives communication from the human shepherd to change their path due to a deviation from the predicted flock behaviour, identifying a need for understanding of the human pilot behaviours. The new direction from the shepherd is based on their understanding of the drone pilot and the response of the sheep.

The levels of desired transparency are set by the shepherding agent, and is based on an agreed semantic map between the shepherd agent and the drone pilot agent, minimal explanations from the drone pilot agent (derived from the semantic map) and inferred by macro and micro-behaviours exhibited by the swarm. Consequently, the communicative transparency in this system is asymmetric and based on the shepherd agent's understanding, with minimal consideration of the drone pilot or a swarm agent context.

The operationalisation of the transparency tenets is described through the cases in Section III-D. The first interaction between the system agents commences with the confusing transparency case, where insufficient interpretability creates a knowledge-gap due to misunderstanding between the shepherd and the pilot. As the agents develop a mature semantic map, their level of interpretability increases. This results in a baseline level of general information exchange that is used as the basis to build from, moving to a case of communicative transparency.

As the cognitive agents within the system gain experience, they refine the language used and employ explanations based on what has been observed, a case of rationally communicative agents, which in turn develops a shared understanding of behaviour and states. When a sufficient level of information symmetry has been obtained between agents, agents develop mutual predictability, switching between the rationally communicative case and the socially communicative case.

The interaction could evolve in multiple directions, where interpretability, explainability and predictability continue to evolve and improve, until the three tenets are mature enough to become fit-for-purpose, resulting in a functionally fit-for-purpose transparency.

As we evolve the system, the case of transparency will change from one case to another. For example, the change of a command from "move to the right quickly" to "proceed 45 degrees to the right at speed 10" by the shepherd to the pilot may be to detail the desired state explicitly. This may not increase the explainability or predictability for the pilot as to why the action is being taken, however, enhances the system interpretability through the refinement of the semantic map. A qualified command that may increase more than one tenet of transparency, such as "proceed 45 degrees to the right at speed 10 in order to move the flock away from the tree line" provides the command refined for interpretability, as well as a more granular intent of "why". This qualified command now allows the pilot to develop goal- and path-planning states while working within the prescriptive constraints issued by the shepherd. Providing a more granular intent of "why" increases the amount of communication between agents and may increase the cognitive load required to support task completion [16].

To support the shepherd agent in future tasks, an HST interface that supports decision making, and is based on HST3-Architecture, would enable the shepherd not only to understand the system but also identify when improvements are required and where. This would be possible due to the fact the HST3-Architecture provides symmetry of information understanding. In doing so, when the swarm is evolving, or human and swarm co-evolving, transparency tenets can support effective communication and collaboration. The HST3-Architecture can improve SA, leading to better decision making by the human. In this situation, the HST3-Architecture can enhance control of the flock through projected influence, as the shepherd is better able to articulate what has occurred within the system, what they are intending on doing, and how they will achieve the desired goals. Using the HST3-Architecture, we provide transparency to the shepherd within the system. This agent can interrogate the drone pilot agent to discover answers such as why are you positioning yourself there? Why are you transitioning into this state? Why are you returning to the base? How will you achieve the (immediate or future) goal?

V. Open Research Questions

The HST literature and proposed HST3-Architecture have identified significant challenges and opportunities for future research proposed for human-swarm teaming. In this section, we will highlight a few of what we have assessed as most pertinent challenges in this area.

The first challenge is related to a few design decisions for the interpretability module. One decision is related to the internal representation and language the swarm use to communicate with each other. This language could be predesigned with a particular lexicon based on a detailed analysis of the possible information that the swarm members need to exchange. However, in situations where the swarm needs to operate in a novel environment, and for a longer period of time, the lexicon, ontology, and language need to be learnt, adapted and allowed to evolve. To design an open-ended language for swarm is challenging, both in terms of our ability to manage the exponential growth in complexity that accompanies such a design, and the difficulty to interpret the continuously evolving swarm language to an external actor, such as a human.

It has been established that a loss of trust is an influential variable with a causal effect for human reliance on a swarm [12]. Moreover, there is a risk for humans on over-relying on a swarm when their trust is based only on the known capabilities of the swarm [13]. In environments where the swarm is evolving, or human and swarm co-evolving, transparency must be an essential element to facilitate effective communication and collaboration. Addressing this research question will ensure that oversight and shared understanding can be maintained during phases of evolution, maintaining higher trust and reliability in a swarm which is otherwise not possible with opaque systems. Nevertheless, the non-stationary nature of the internal language within each of these actors due to its evolving abilities will create significant complexity in interpreting the language to external actors, who could also be evolving their own language. It is hard to conceive how to overcome this challenge without allowing heavy communications to occur between the swarm and the human to exchange changes in their lexicons, syntax and semantics.

A main explainability challenge in a swarm is the decentralised nature of reasoning. In a homogeneous swarm, the reasoning process within each agent are the same. While the agents may accumulate different experiences due to them encountering different states in the environment, thus, they may be holding heterogenous knowledge, over a larger operational time-horizon, it would be expected that they converge on similar knowledge. Nevertheless, the human is not observing necessarily every member of the swarm. Instead, the human is observing some or all members simultaneously and needs the aggregate causal chain that led a swarm to reach a particular state or perform a particular set of collective actions. The shepherding research offers a mechanism to overcome this challenge by making the requirement of explanation the responsibility of a few members of agents (the sheepdogs).

A number of previous studies identify meta-categories [108], measures [109], and tenets for consideration [107] to evaluate system transparency. A broad architecture at the system level remains yet to be developed with the ability to increase context-dependent SA through enhanced transparency. However, little research is available that investigates the success or failure of a swarm's transparency. The HST3-architecture offers a design where more research could be conducted on the individual tenets of transparency and to isolate the effects of each tenet on system performance and agent's trust.

VI. Conclusion

We have proposed a portfolio of definitions for the vital concept of transparency, and its tenets interpretability, explainability, and predictability, within the setting of human-swarm teaming (HST). These measures describe these constituent elements, and their contribution to designing transparency in HST settings, essential elements for human trust. Our work addresses the need within the literature to clearly define these terms and present cases that differentiate how they are used. The proposed architecture answers the question of "how" to develop transparency in HST, providing a systems approach to enabling SAT. Within HST3-Architecture, reliability can be measured and evolved by leveraging the tenet of predictability.

Our architecture has general applicability, particularly in situations where a shared understanding is required, to help practitioners and researchers realise transparency for HST. Example fields for application include security and emergency services where operational assurance and decision traceability are required, or as importantly agricultural settings where tasks may be outsourced to increase productivity.

VII. Acknowledgement

Abbass's contribution is funded by a USA Office of Naval Research - Global (ONR-G) Grant No. N629091812140.

References

[1] Z. Chen, H. Lu, S. Tian, J. Qiu, T. Kamiya, S. Serikawa, and L. Xu, "Construction of a hierarchical feature enhancement network and its application in fault recognition," IEEE Transactions on Industrial Informatics, 2020.

[2] P. Wang, D. Wang, X. Zhang, X. Li, T. Peng, H. Lu, and X. Tian, "Numerical and experimental study on the maneuverability of an active propeller control based wave glider," Applied Ocean Research, vol. 104, p. 102369, 2020.

[3] G. Beni, "From swarm intelligence to swarm robotics," in International Workshop on Swarm Robotics. Springer, 2004, pp. 1–9.

[4] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: a review from the swarm engineering perspective," Swarm Intelligence, vol. 7, no. 1, pp. 1–41, 2013.

[5] P. Zhu, W. Dai, W. Yao, J. Ma, Z. Zeng, and H. Lu, "Multi-robot flocking control based on deep reinforcement learning," IEEE Access, vol. 8, pp. 150 397–150 406, 2020.

[6] E. Şahin, "Swarm robotics: From sources of inspiration to domains of application," in International workshop on swarm robotics. Springer, 2004, pp. 10–20.

[7] W. D. Nothwang, M. J. McCourt, R. M. Robinson, S. A. Burden, and J. W. Curtis, "The human should be part of the control loop?" in 2016 Resilience Week (RWS). IEEE, Aug 2016.

[8] A. Kolling, S. Nunnally, and M. Lewis, "Towards human control of robot swarms," in Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction. ACM, 2012, pp. 89–96.

[9] A. Hussein and H. Abbass, "Mixed initiative systems for human-swarm interaction: Opportunities and challenges," in 2018 2nd Annual Systems Modelling Conference (SMC). IEEE, 2018, pp. 1–8.

[10] H. A. Abbass, "Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust," Cognitive Computation, vol. 11, no. 2, pp. 159–171, 2019.

[11] H. A. Abbass, E. Petraki, K. Merrick, J. Harvey, and M. Barlow, "Trusted autonomy and cognitive cyber symbiosis: Open challenges," Cognitive Computation, vol. 8, no. 3, pp. 385–408, 2016.

[12] A. Hussein, S. Elsawah, and H. A. Abbass, "Trust mediating reliability–reliance relationship in supervisory control of human–swarm interactions," Human Factors, 2019, pMID: 31590574.

[13] ——, "The reliability and transparency bases of trust in human-swarm interaction: principles and implications," Ergonomics, no. just-accepted, pp. 1–19, 2020.

[14] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004.

[15] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," Ergonomics, vol. 35, no. 10, pp. 1243–1270, 1992.

[16] M. R. Endsley, "From Here to Autonomy: Lessons Learned From Human-Automation Research," Human Factors, vol. 59, no. 1, pp. 5–27, 2017.

[17] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci, "Intelligent agent transparency in human–agent teaming for multi-uxv management," Human Factors, vol. 58, no. 3, pp. 401–415, 2016.

[18] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," Proceedings of the National Academy of Sciences, vol. 116, no. 44, pp. 22 071–22 080, Oct. 2019, 00025 Publisher: National Academy of Sciences Section: Physical Sciences.

[19] S. Hara and K. Hayashi, "Making tree ensembles interpretable," arXiv preprint arXiv:1606.05390, 2016.

[20] E. Debie, K. Shafi, C. Lokan, and K. Merrick, "Reduct based ensemble of learning classifier system for real-valued classification problems," in 2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL). IEEE, 2013, pp. 66–73.

[21] E. Debie, K. Shafi, K. Merrick, and C. Lokan, "An online evolutionary rule learning algorithm with incremental attribute discretization," in 2014 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2014, pp. 1116–1123.

[22] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," Machine Learning, vol. 102, no. 3, pp. 349–391, 2016.

[23] A. Jobin, I. Marcello, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.

[24] M. Turilli and L. Floridi, "The ethics of information transparency," Ethics and Information Technology, vol. 11, no. 2, pp. 105–112, 2009.

[25] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," Army research lab aberdeen proving ground md human research and engineering, Tech. Rep., 2014.

[26] K. A. Roundtree, M. A. Goodrich, and J. A. Adams, "Transparency: Transitioning from human–machine systems to human-swarm systems," Journal of Cognitive Engineering and Decision Making, vol. 13, no. 3, pp. 171–195, September 2019.

[27] J. A. Adams, J. Y. Chen, and M. A. Goodrich, "Swarm transparency," in Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 2018, pp. 45–46.

[28] R. Liu, F. Jia, W. Luo, M. Chandarana, C. Nam, M. Lewis, and K. Sycara, “Trust-aware behavior reflection for robot swarm self-healing,” in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 122–130.

[29] S. Nunnally, P. Walker, A. Kolling, N. Chakraborty, M. Lewis, K. Sycara, and M. Goodrich, “Human influence of robotic swarms with bandwidth and localization issues,” in 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct 2012, pp. 333–338.

[30] K. A. Roundtree, M. D. Manning, and J. A. Adams, “Analysis of human-swarm visualizations,” in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 62, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 287–291.

[31] M. Schranz, G. A. Di Caro, T. Schmickl, W. Elmenreich, F. Arvin, A. Şekercioğlu, and M. Sende, “Swarm intelligence and cyber-physical systems: Concepts, challenges and future trends,” Swarm and Evolutionary Computation, vol. 60, 2021.

[32] A. Bagnato, R. Bíró, D. Bonino, C. Pastrone, W. Elmenreich, R. Reiners, M. Schranz, and E. Arnautovic, “Designing swarms of cyber-physical systems: the h2020 cpswarm project: Invited paper,” in Proceedings of the Computing Frontiers Conference. Association for Computing Machinery, 2017, pp. 305–312.

[33] A. Hussein, S. Elsawah, and H. A. Abbass, “Swarm collective wisdom: a fuzzy-based consensus approach for evaluating agents confidence in global states,” in The IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2020.

[34] U. Premarathne and S. Rajasingham, “Trust based multi-agent cooperative load balancing system (tclbs),” Future Generation Computer Systems, 2020.

[35] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” Human Factors, vol. 37, no. 1, pp. 32–64, 1995.

[36] J. Y. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, “Situation awareness-based agent transparency and human-autonomy teaming effectiveness,” Theoretical Issues in Ergonomics Science, vol. 19, no. 3, pp. 259–282, 2018.

[37] J. L. Wright, J. Y. Chen, and S. G. Lakhmani, “Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability,” IEEE Transactions on Human-Machine Systems, vol. 50, no. 3, pp. 254–263, 2020.

[38] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Interpretable machine learning: definitions, methods, and applications,” arXiv preprint arXiv:1901.04592, 2019, 00052.

[39] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019, 00514.

[40] B. Kim, R. Khanna, and O. O. Koyejo, “Examples are not enough, learn to criticize! criticism for interpretability,” in Advances in neural information processing systems, 2016, pp. 2280–2288, 00164.

[41] O. Biran and C. Cotton, “Explanation and justification in machine learning: A survey,” in IJCAI-17 workshop on explainable AI (XAI), vol. 8, 2017, p. 1, 00088.

[42] M. Sierhuis and S. J. B. Shum, “Human-agent knowledge cartography for e-science: NASA field trials at the Mars Desert Research Station,” in Knowledge cartography. Springer, 2014, pp. 381–399, 00012.

[43] A. Lazaridou, A. Peysakhovich, and M. Baroni, “Multi-agent cooperation and the emergence of (natural) language,” arXiv preprint arXiv:1612.07182, 2016, 00166.

[44] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra, “Learning Cooperative Visual Dialog Agents With Deep Reinforcement Learning,” in ICCV 2017, 2017, pp. 2951–2960, 00237.

[45] J. Andreas, A. Dragan, and D. Klein, “Translating neuralese,” arXiv preprint arXiv:1704.06960, 2017, 00028.

[46] D. St-Onge, F. Levillain, E. Zibetti, and G. Beltrame, “Collective expression: how robotic swarms convey information with group motion,” Paladyn, Journal of Behavioral Robotics, vol. 10, no. 1, pp. 418–435, 2019, 00001 Publisher: De Gruyter.

[47] A. Suresh and S. Martínez, “Human-swarm Interactions for Formation Control Using Interpreters,” International Journal of Control, Automation and Systems, pp. 1–14, 2020, 00000 Publisher: Springer.

[48] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati, “Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior,” in Proceedings of the International Conference on Automated Planning and Scheduling, vol. 29, 2019, pp. 86–96, 00027 Issue: 1.

[49] M. Fox, D. Long, and D. Magazzeni, “Explainable planning,” in IJCAI-17 workshop on Explainable AI, 2017.

[50] W. R. Swartout and J. D. Moore, “Explanation in second generation expert systems,” Second Generation Expert Systems, 1993.

[51] M. Minksy, The Society of Mind. Simon and Schuster, 1985.

[52] D. Doran, S. Schulz, and T. Besold, “What does explainable ai really mean? a new conceptualization of perspectives,” in Comprehensibility and Explanation in AI and ML, 2017. [Online]. Available: http://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf

[53] M. van Lent, W. Fisher, and M. Mancuso, “An explainable artificial intelligence system for small-unit tactical behavior,” American Association for Artificial Intelligence: IAAI Emerging Applications, 2004.

[54] J. Josephson and S. Josephson, Abductive Inference: Computation, Philosophy, Technology. Cambridge University Press, 1996.

[55] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, 2019.

[56] M. Turek, “Explainable artificial intelligence (xai),” https://www.darpa.mil/program/explainable-artificial-intelligence, April 2020.

[57] E. Diamant, “Designing artificial cognitive architectures: Brain inspired or biologically inspired?” Procedia Computer Science, vol. 145, pp. 153 – 157, 2018, postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society), held August 22-24, 2018 in Prague, Czech Republic.

[58] F. Dosilovic, M. Brcic, and N. Hlupic, “Explainable artificial intelligence: A survey,” in MIPRO 2018 - 41st International Convention Proceedings, 2018.

[59] E. Nyamsuren and N. A. Taatgen, “Human reasoning module,” Biologically Inspired Cognitive Architectures, vol. 8, pp. 1 – 18, 2014.

[60] D. Amir and O. Amir, “Highlights: Summarizing agent behavior to people,” in Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, ser. AAMAS ’18. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 1168–1176.

[61] J. van der Waa, J. van Diggelen, K. van den Bosch, and M. Neerincx, “Contrastive explanations for reinforcement learning in terms of expected consequences,” 2018.

[62] W. L. Johnson, “Agents that learn to explain themselves,” AAAI-94 Proceedings, 1994.

[63] J. R. Searle, “Minds, brains, and programs,” Behavioral and Brain Sciences, vol. 3, no. 3, p. 417–424, 1980.

[64] A. Turing, “Computing machinery and intelligence,” Mind, pp. 433–460, October 1950.

[65] A. Rosenfeld and A. Richardson, “Explainability in human-agent systems,” Autonomous Agents and Multi-Agent Systems, pp. 1–33, 2019.

[66] R. Hoffman, G. Klein, and S. Mueller, “Explaining explanation for "explainable ai",” in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 1, 2018, pp. 197–201.

[67] J. Lyons, “Being transparent about transparency: A model for human-robot interaction,” in AAAI Spring Symposium - Technical Report, vol. SS-13-07, 2013, pp. 48–53.

[68] C. Nam, H. Li, S. Li, M. Lewis, and K. Sycara, “Trust of humans in supervisory control of swarm robots with varied levels of autonomy,” in 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 10 2018, pp. 825–830.

[69] Z. Gong and Y. Zhang, “Behavior explanation as intention signaling in human-robot teaming,” in RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication, 2018, pp. 1005–1011.

[70] J. Penders, L. Alboul, U. Witkowski, A. Naghsh, J. Saez-Pons, S. Herbrechtsmeier, and M. El-Habbal, “A robot swarm assisting a human fire-fighter,” Advanced Robotics, vol. 25:1-2, pp. 93–117, 2011.

[71] C. University, predict. Cambridge University Press, 2020. [Online]. Available: https://dictionary.cambridge.org/dictionary/essential-british-english/predict

[72] E. J. Stanek Iii, J. da Motta Singer, and V. B. Lencina, “A unified approach to estimation and prediction under simple random sampling,” Journal of Statistical Planning and Inference, vol. 121, no. 2, pp. 325–338, 2004.

[73] N. B. Masese, G. M. Muketha, and S. M. Mbuguah, “Interface features, program complexity and memorability as indicators of learnability of mobile social software,” International Journal of Science and Research (IJSR), 2017.

[74] S. Dimitriadis, A. Kouremenos, and N. Kyrezis, “Trust-based segmentation,” International Journal of Bank Marketing, 2011.

[75] S. Behzadi, B. Schelling, and C. Plant, “Itgh: Information-theoretic granger causal inference on heterogeneous data,” in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2020, pp. 742–755.

[76] B. M. Muir, “Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems,” Ergonomics, vol. 37, no. 11, pp. 1905–1922, 1994.

[77] S. M. Merritt and D. R. Ilgen, “Not all trust is created equal: Dispositional and history-based trust in human-automation interactions,” Human Factors, vol. 50, no. 2, pp. 194–210, 2008.

[78] K. Drnec and J. S. Metcalfe, “Paradigm development for identifying and validating indicators of trust in automation in the operational environment of human automation integration,” in International Conference on Augmented Cognition. Springer, 2016, pp. 157–167.

[79] G. Klien, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, “Ten challenges for making automation a" team player" in joint human-agent activity,” IEEE Intelligent Systems, vol. 19, no. 6, pp. 91–95, 2004.

[80] L. L. Constantine, “Trusted interaction: User control and system responsibilities in interaction design for information systems,” in International Conference on Advanced Information Systems Engineering. Springer, 2006, pp. 20–30.

[81] R. Marín, P. J. Sanz, and A. P. Del Pobil, “The uji online robot: An education and training experience,” Autonomous Robots, vol. 15, no. 3, pp. 283–297, 2003.

[82] R. C. Winck, S. M. Sketch, E. W. Hawkes, D. L. Christensen, H. Jiang, M. R. Cutkosky, and A. M. Okamura, “Time-delayed teleoperation for interaction with moving objects in space,” in 2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014, pp. 5952–5958.

[83] C. Sherstan, J. Modayil, and P. M. Pilarski, “A collaborative approach to the simultaneous multi-joint control of a prosthetic arm,” in 2015 IEEE International Conference on Rehabilitation Robotics (ICORR). IEEE, 2015, pp. 13–18.

[84] L. Ding, H. Gao, Z. Deng, Y. Li, G. Liu, H. Yang, and H. Yu, “Three-layer intelligence of planetary exploration wheeled mobile robots: Robint, virtint, and humint,” Science China Technological Sciences, vol. 58, no. 8, pp. 1299–1317, 2015.

[85] M. Panzirsch, H. Singh, M. Stelzer, M. J. Schuster, C. Ott, and M. Ferre, “Extended predictive model-mediated teleoperation of mobile robots through multilateral control,” in 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 1723–1730.

[86] K. Yerex, D. Cobzas, and M. Jagersand, “Predictive display models for tele-manipulation from uncalibrated camera-capture of scene geometry and appearance,” in 2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422), vol. 2. IEEE, 2003, pp. 2812–2817.

[87] P. Walker, A. Kolling, S. Nunnally, N. Chakraborty, M. Lewis, and K. Sycara, “Investigating neglect benevolence and communication latency during human-swarm interaction,” in 2012 AAAI Fall Symposium Series, 2012.

[88] J. C. Lane, C. R. Carignan, and D. L. Akin, “Time delay and communication bandwidth limitation on telerobotic control,” in Mobile Robots XV and Telemanipulator and Telepresence Technologies VII, vol. 4195. International Society for Optics and Photonics, 2001, pp. 405–419.

[89] J.-W. Park, C.-D. Kim, and J.-M. Lee, “Concurrent bilateral teleoperation over the internet,” in ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No. 01TH8570), vol. 1. IEEE, 2001, pp. 302–307.

[90] D. Q. Huy, I. Vietcheslav, and G. S. G. Lee, “See-through and spatial augmented reality-a novel framework for human-robot interaction,”

[93] M. E. Walker, H. Hedayati, and D. Szafir, “Robot teleoperation with augmented reality virtual surrogates,” in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2019, pp. 202–210.

in 2017 3rd International Conference on Control, Automation and Robotics (ICCAR). IEEE, 2017, pp. 719–726.

[91] M. T. Rosenstein, A. H. Fagg, S. Ou, and R. A. Grupen, “User intentions funneled through a human-robot interface,” in Proceedings of the 10th international conference on Intelligent user interfaces, 2005, pp. 257–259.

[92] A. Zignoli, F. Biral, K. Yokoyama, and T. Shimono, “Including a musculoskeletal model in the control loop of an assistive robot for the design of optimal target forces,” in IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society, vol. 1. IEEE, 2019, pp. 5394–5400.

[94] A. Hermann, F. Mauch, K. Fischnaller, S. Klemm, A. Roennau, and R. Dillmann, “Anticipate your surroundings: Predictive collision detection between dynamic obstacles and planned robot trajectories on the gpu,” in 2015 European Conference on Mobile Robots (ECMR). IEEE, 2015, pp. 1–8.

[95] J. Fisac, A. Bajcsy, S. Herbert, D. Fridovich-Keil, S. Wang, C. Tomlin, and A. Dragan, “Probabilistically safe robot planning with confidence-based human predictions,” in Proceedings of Robotics: Science and Systems, Pittsburgh, Pennsylvania, June 2018.

[96] P. Walker, S. Nunnally, M. Lewis, A. Kolling, N. Chakraborty, and K. Sycara, “Neglect benevolence in human-swarm interaction with communication latency,” in International Conference on Swarm, Evolutionary, and Memetic Computing. Springer, 2012, pp. 662–669.

[97] ——, “Neglect benevolence in human control of swarms in the presence of latency,” in national Conference on Systems, Man, and Cybernetics. IEEE, 2012, pp. 3009–3014.

[98] Z. Lipton, “The mythos of model interpretability,” Proc. ICML Workshop Hum. Interpretability Mach. Learn., pp. 96–100, 2016.

[99] S. B. Hale, Overview of the Field of Interpreting and Main Theoretical Concepts. London: Palgrave Macmillan, 2007.

[100] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, “Sentic lstm: a hybrid network for targeted aspect-based sentiment analysis,” Cognitive Computation, vol. 10, pp. 639–650, 2018.

[101] X. Sun, X. Peng, and S. Ding, “Emotional human-machine conversation generation based on long short-term memory,” Cognitive Computation, vol. 10, pp. 389–397, 2018.

[102] M. Friedman, “Explanation and scientific understanding,” The Journal of Philosophy, vol. 71, no. 1, pp. 5–19, 1974.

[103] P. Johnson-Laird and M. Ragni, “Possibilities as the foundation of reasoning,” Cognition, vol. 193, p. 103950, 2019.

[104] A. Anderson, J. Dodge, A. Sadarangani, Z. Juozapaitis, E. Newman, J. Irvine, S. Chattopadhyay, M. Olson, A. Fern, and M. Burnett, “Mental models of mere mortals with explanations of reinforcement learning,” ACM Trans. Interact. Intell. Syst., vol. 10, no. 2, May 2020.

[105] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” 2019.

[106] P. Langley, B. Meadows, M. Sridharan, and D. Choi, “Explainable agency for intelligent autonomous systems,” in AAAI, 2017.

[107] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, 2019.

[108] S. Mohseni, N. Zarei, and E. D. Ragan, “A multidisciplinary survey and framework for design and evaluation of explainable ai systems.” arXiv: Human-Computer Interaction, 2020.

[109] A. as an Intelligent Teammate: Social Psychological Implications, “Jose kerstholt and jonathan barnhoorn and tom hueting and lotte schuilenborg,” NATO S&T Organization, Tech. Rep. STO-MP-HFM-300, nd.

[110] P. Walker, M. Lewis, and S. K, “The effect of display type on operator prediction of future swarm states,” in 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, 2016, pp. 002 521–002 526.