Operational Trust in Mission Autonomy (OPTIMA) Final Report

Dr. Cara LaPointe, Dr. Sarah Rigsbee, Ms. Alexis Basantis, Mr. Bryan Camacho, Mr. Matthew Tomaszewski

September 2023

Prepared for: Office of the Undersecretary of Defense for Research and Engineering
Prepared by: The Johns Hopkins University (JHU) Applied Physics Laboratory (APL) and The Johns Hopkins University Institute for Assured Autonomy (IAA)

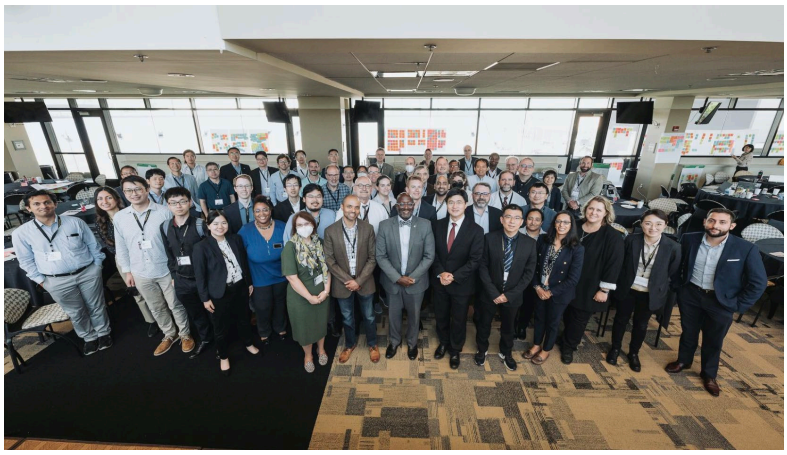
Contents

Executive Summary

Autonomous systems are becoming increasingly more complex and more ubiquitous across a variety of domains. Specifically in complex, operational scenarios, these systems have the potential to have a large impact, primarily aiding the humans living and working in these environments. Therefore, understanding operational trust in mission autonomy is critical for the appropriate and responsible design, development, implementation, and use of complex, intelligent systems.

The following report outlines work done for the Operational Trust in Mission Autonomy (OPTIMA) initiative by a team of design strategists at the Johns Hopkins Institute for Assured Autonomy (IAA). The charter of this initiative states that understanding operational trust is incredibly important because it will be key for "delivering trusted Autonomy for Robotic and Intelligent Autonomous Systems (RAS) in complex, contested missions on the multi-domain battlefield, and enabling deployment of effective human-machine teams." In addition, it states that there is a "lack of existing frameworks within the Department of Defense (DoD) to define quantifiable metrics for trust, and a lack of uniform concepts for architecting trust into RAS", clearly articulating a key gap that must be addressed in order to make trusted battlefield autonomy a reality. Therefore, our team leveraged a human-centered design and design thinking approach to examine the question: "how might we identify and/or develop metrics of operational trust in mission autonomy?"

In order to further investigate this question, the IAA team, in collaboration with Purdue University's Center for Institute for Control, Optimization, and Networks (ICON), and the National Security Innovation Network (NSIN), planned and facilitated a Summit on Trusted Autonomy Research and Technology (START) which took place in July 2022 at Purdue University. The goal of this event was to convene and equitably solicit opinions from a diverse set of stakeholders and experts from academia, DoD, and commercialization sectors to define a multi-year research and development (R&D) and investment roadmap to enable trusted autonomy for OPTIMA and for broader societal applications. The summit hosted a total of 118 participants from 63 unique sector entities, representing academia, DoD, venture capitalists, industry, and other organizations, such as Federally Funded Research and Development Centers (FFRDCs). During the summit, these participants collaborated together and brainstormed different metrics, methods, and investment opportunities around human-centric, machine-centric, and integration-centric topics.



OPTIMA START participant group photo

Across all participant groups, over 2000 qualitative datapoints were captured throughout the generative brainstorming sessions. This qualitative data was digitized, organized, and synthesized by the IAA team into 16 specific metrics, spanning over 5 thematic areas, which outline both qualitative and quantitative metrics for measuring mission autonomy trust impact factors. In addition, the synthesized data yielded 88 design and implementation principles, spanning over 10 thematic areas, which aim to act as conversation starters or guidelines for designers, developers, builders, etc. of intelligent, autonomous systems.

START brainstorming action shots

Our team recognizes that although the metrics and guiding principles from the data are a great starting point and provide a representation of the voices and perspectives that participated in

the event, they are not all encompassing. Future work must be done to further investigate these topics to better understand their nuances and potential applications.

In November 2022, a convening was held at the Johns Hopkins Applied Physics Lab (JHU/APL) with key industry, government, and academic partners. The goal of this meeting was to develop connections and assemble the objectives, progress, activities, and path forward for each of the key players/efforts across the OPTIMA initiative. These efforts were then mapped to key objectives of the initiative and potential gaps were identified. All of this work was consolidated and culminated into the OPTIMA Playbook, which was distributed and briefed out to organization leadership. Additional OPTIMA Playbook meeting details and initial findings can be found in Appendix E.

In addition to generating a final report, portions of the work described about have been shared with international Allies and partners. In July 2023, the team gave a presentation at the First International Symposium on Trustworthy Autonomous Systems (TAS '23) focused on the HCD method of developing and facilitation the OPTIMA workshop. As well as giving an oral presentation, the work was also published as a peer-reviewed paper in the TAS '23 Symposium Proceedings.

Dr. Sarah Rigsbee presenting the HCD OPTIMA methods at TAS '23

Because the dataset derived from OPTIMA was so robust and complex, we hope to have the opportunity to continue to perform data analysis and publish additional peer-reviewed papers, expanded to include the nuanced insights and findings found in the data, in additional highly-regarded journals and conferences. Additional areas of research and exploration include (but are not limited to):

- Robust literature reviews on metric themes derived from START
- Diving deeper into the metrics derived from START through additional data synthesis and workshops
- Exploring trust in autonomy within different communities or populations (e.g., different military domains, different theaters) through human-centered design and design thinking processes

Introduction

As technology continues to evolve, advanced systems are both becoming more complex and more integrated into our lives and the world. As these systems advance and move from "automatic" to "automated" to "autonomous," human perspectives of these systems tend to shift from thinking of them as "tools" to thinking of them as "partners" [1]. From smart devices that can help order groceries to advanced AI-enabled algorithms that can aid in battlefield decision making, these systems can provide assistance to human users on a variety of scales and in an array of domains. Regardless of task, autonomous or semi-autonomous systems are becoming crucial partners to human users. Therefore, the concept of trust in these systems is also becoming increasingly important [2].

Trust is a large, complex, and multilayered concept to explore. It is also critical to achieving effective and willing human-machine relationships. We recognize that human trust is paramount for the relationship of humans and machines to be effective, and this is true not only at the individual user perspective but also needs to account for the larger group or community perspectives of trust.

This paper presents general research on human-centered design and trust in autonomy. However, its specific aim is to outline a method, which leveraged and adapted existing human-centered design approaches, to explore a novel and challenging problem set: investigating metrics of human operational trust in autonomy, grounded in the social, cultural, and ecological contexts in which the autonomous system will ultimately operate. In addition to outlining human-centered design techniques, this paper provides an example of a proactive approach for innovators, researchers, and stakeholders to explore challenges in ethically and socially conscious ways, especially when it comes to how autonomy is developed and integrated into contested military environments. Since this paper's primary focus is on the description and development of these techniques and this method, preliminary results presented herein are merely for context and example. Additional analysis and synthesis of data are forthcoming.

The Challenge

Our research aimed to better understand the nature of trust in systems used in tactical, operational environments and to explore the factors that influence trust in these high-stress settings. Therefore, we partnered with the U.S. Office of the Secretary for Defense (OSD) and the Operational Trust in Mission Autonomy (OPTIMA) initiative to examine the question: "how might we identify and/or develop metrics of operational trust in mission autonomy?"

The OPTIMA initiative states that understanding operational trust is incredibly important because it will be key for "delivering trusted Autonomy for Robotic and Intelligent Autonomous Systems (RAS) in complex, contested missions on the multi-domain battlefield, and enabling deployment of effective human-machine teams [3]." In addition, it states that there is a "lack of existing frameworks within the Department of Defense (DoD) to define quantifiable metrics for trust, and a lack of uniform concepts for architecting trust into RAS", clearly articulating a key gap that must be addressed in order to make trusted battlefield autonomy a reality [3].

Trust

The concept of trust has been examined for millennia and published research can be found in almost every single academic domain – from engineering to philosophy. However, trust, in its most simple and generic definition, is the "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party" [4]. As noted, trust is rooted in an individual's expectations and perceptions of the relationship and can be affected by a multitude of different aspects (e.g., reliability, honesty, motivations, etc.) [5]. Although these definitions and frameworks of trust are typically focused on human relationships, as the integration of technology into daily life is increasing, they have also started to be applied to human-machine relationships.

When considering trust in technology, the literature defines it as beliefs or attitudes that a specific technology has the attributes necessary to perform as expected and contribute to a goal in a given situation in which negative consequences are possible (e.g., vulnerabilities or uncertainty) [6, 7]. Similar to human relationships, trust and reliance are both strongly linked to expectations of behaviors and performance of both the humans and the machine [1, 8]. In addition, it has been found that increased familiarity and transparency of the machine can facilitate the development and maintenance of trust [9].

As is seen in literature, media, and our own lives, trust in technology during a "typical" day-to-day scenario is difficult to understand, engender, and maintain. When thinking about environments that are even higher-stress, higher-risk, and more complex, such as in military domains and operational scenarios, trust is even more difficult to understand and measure. Therefore, our efforts specifically focused on this concept of "operational trust" and how it relates to mission autonomy.

Operational Trust

As mentioned previously, across a variety of focus areas and domains, a large body of research exists on the general construct of trust (e.g., how humans trust others) and human-machine teaming (e.g., how to humans interact with machines to achieve a goal). Although we recognize that operational trust is related to the larger research area of human-machine teaming, studies that specifically call out "operational trust" in the context of autonomous systems are sparse [10, 11]. Only a handful of references exist that clearly articulate definitions of operational trust in these contexts [12, 13]. According to these sources, at its core, operational trust must consider all users in the ecosystem (e.g., warfighters, peers, commanders, subordinates) and how the automated systems meet the users' performance expectations [13]. These systems can be leveraged for knowledge sharing and workload division in order to accomplish complex mission operations [12].

Narrowing the focus to operational trust from the broader concepts of trust and human-machine teaming enabled more targeted efforts to examine mission-specific environments, scenarios, and user types that contribute to this construct. By better understanding these

components, it lays the foundation for investigating potential metrics of operational trust which could contribute to better system design, adoption, and use in these domains.

Leveraging a Human-Centered Design Approach

Even when working in human-machine teams, trust is directly influenced and shaped by the human users' perspectives, expectations, and emotions [5]. It is not enough for systems to be trustworthy, a concept that is more technology-focused, where the system acts reliably, safely, and degrades gracefully. These systems must also be trusted, a more human-centered concept, which focuses on the confidence that the user has with a particular system [14]. Therefore, it is important to leverage human minds to begin to understand, unpack, and design for system trustworthiness and user trust.

Design is a ubiquitous process that surrounds everyday life: it is what connects products, services, environments, and people and has the power to shape and improve human experiences. However, design problems cannot be solved in a vacuum, especially in a world where the interconnection of people, places, and things is rapidly growing [15]. This increase in connection has emphasized the importance of representing these relationships and perspectives in the understanding and solving of design challenges. Integrating the human perspective into problem solving can take on a variety of different forms, however two commonly referred to approaches are User Centered Design (UCD) and Human Centered Design (HCD).

UCD is an approach which focuses specifically on how end users interact with technology [16], thus applying a specific end user group lens to investigating challenge spaces. This approach can work well in addressing the needs of the specific user, for a specific and predetermined tool that is being developed [17]. The UCD process facilitates direct input and feedback from the individuals using the product on the way it is designed and implemented. In our experience, we have found that UCD problems are typically relatively constrained, for example, developing a graphical user interface (GUI) on a singular system which serves a well-defined end user group or demographic. By this inherently narrow focus, UCD methods can have fixed and/or preconceived notions [18] on the "correct" solution space and may fail to promote broader human interests [16].

Whereas UCD focuses on a well-defined user group, HCD shifts the focus to the broader role of people [19] and also acknowledges that people's hopes, fears, dreams and aspirations will affect the way that technology is responded to and interacted with [20]. This type of design can be achieved through leveraging a top-down (driven by decision makers and experts), bottom-up (driven by local users), or a hybrid (combination of both) approach [21]. By leveraging this type of hybrid approach, the perspectives of all stakeholders are considered - not only the end users but also other individuals or groups that might be impacted by or involved in the challenge.

HCD approaches lend themselves well when "wicked problems" are being tackled [22, 23]. Wicked problems are a class of ill formulated social problems, with confusing information and where many users and decision-makers may have conflicting values [24]. The nature of these

wicked problems requires the human perspective to recognize the complexity of the problem, understand and make sense of the subsequent challenges, and then make them manageable [25].

For example, when dealing with complex technical systems, there are often a multitude of interconnected factors that can impact the effectiveness and usability of that system. Navigating such complex landscapes and identifying appropriate approaches can be a challenge [26]. HCD approaches to such complex systems can include investigating questions from the physical nature of human-product interactions (who), the activities, tasks and functions (what), what interactivity may look like (when), communication and discourse (how), all the way to the metaphysical meaning (why) [27].

While there are inherent challenges and tensions to consider when applying HCD, through critical reflection and thoughtful approaches it can be an incredibly powerful approach to leverage [19]. Engaging with a variety of individuals, rather than an individual end user group, that the system affects helps designers gain a better understanding of these connection points and the broader social, cultural, and ecological contexts in which the system or environment operates [22]. This more holistic understanding can lead to a more thoughtful and impactful design solution, which considers the entire system and operating environment.

In summary, both approaches leverage similar processes and prioritize empathy and co-design as essential problem-solving pillars. Both are iterative and participatory in nature of solution development, leading to concepts that are not only effective for the problem at hand, but also sustainable and equitable in an ever-changing, dynamic context. The key distinction between them is the user group that is considered. UCD focuses more narrowly on a specific end user base, whereas HCD methods approach a challenge from a more holistic perspective, accounting for multiple stakeholders such as end users, designers, developers, integrators, and decision-makers.

The challenge we set forth to investigate in this paper around metrics of operational trust is ambiguous and complex. It has never been thoroughly examined before and there is no specific end user group that can provide concrete guidance or answers. It will take a community of individuals, including warfighters, academic researchers, engineers, legislators, and countless others to begin to tackle this challenge, leveraging HCD techniques that communicate, interact, empathize and engage with the people involved [17].

Methods

Although the methods described above can be applied to tackle a variety of ambiguous challenges, our team applied HCD methodologies specifically to the large, ambiguous challenge of developing and identifying metrics of operational trust in mission autonomous systems. Metrics, by definition, are measures that are quantifiable and can represent both objective (undeniable facts) and subjective (based on personal judgment) data [28, 29].

To break down the vast challenge space, the team leveraged a framework developed by the Johns Hopkins Institute for Assured Autonomy (IAA). The IAA is "a national center of excellence ensuring the safe, secure, reliable, and predictable integration of autonomous systems into society by covering the full spectrum of research across the three pillars of technology, ecosystem, and ethics and governance" whose goal is to "ensure that autonomous systems will be trusted to operate as expected, to respond safely to unexpected inputs, to withstand corruption by adversaries, and to integrate seamlessly into society [14]." The team leveraged the Johns Hopkins Human-Centered Trust (JHHT) Framework, which was developed to identify the key stakeholder groups who should be represented in the design, integration, and use of complex, automated systems. These stakeholder groups include: individual users, communities, institutions, and the systems themselves [30]. For this effort, to better understand operational trust in technical systems and begin to identify metrics of trust, we focused on engaging user communities and governing / building institutions. User communities, which included individuals with current or prior operational experience/military service, offered perspectives around trust in operational scenarios and previous experience with technical systems in these environments. Represented governing and building institutions, such as the DoD, academia, and laboratories, brought background and perspectives focused on system performance, education, and ethical metrics.

Method Development

Human-machine teaming and trusted autonomy will be enduring and critical across military and civilian domains and within a multitude of near and far term applications. To ensure these perspectives are represented, the IAA, the Purdue University's Center for Institute for Control, Optimization, and Networks (ICON), and the National Security Innovation Network (NSIN) collaborated together to specifically design an ICON hosted summit to facilitate robust conversations that will shape requirements for future DoD investment on research and autonomy. ICON focuses on integrating diverse expertise to provide innovative solutions to grand challenges in automation, with a higher-level mission of bringing together academic researchers to collaborate with industry, government agencies, and national labs to tackle problems of national and global priority in autonomy [31]. NSIN is a unique problem-solving network within the DoD, adapting to the emerging needs of those who serve in the defense of our national security through bringing together defense, academic, and entrepreneurial innovations to solve national security challenges [32]. The planning committee for this event was composed of IAA, ICON, and NSIN individuals that spanned across multiple technical (e.g., multi-disciplinary engineering domains) and non-technical (e.g., business development) backgrounds, working professions, areas of expertise, ethnicities, gender expression and identities, organizations, and sectors (government, academia, and not-for-profit university affiliated research center). The depth and breadth of diversity on the planning committee helped identify opportunities for strengthening inclusion of different participant types in the most equitable way possible within the constraints of the event. The Summit on Trusted Autonomy Research and Technology (START) was thus designed and developed to convene and equitably solicit opinions from a diverse set of stakeholders and experts from academia, DoD, and commercialization sectors to define a multi-year research and development (R&D) and

investment roadmap to enable trusted autonomy for OPTIMA and for broader societal applications.

Working closely with our collaborators from ICON and NSIN, the objective to explore trusted autonomy more holistically across the human-autonomy team was guided by focusing on the human, machine, and human-machine integrative perspectives. Exploring these three facets of operational trust helped define and explore what might be a roadmap for enabling trusted autonomy. To that end, the summit focused on convening a diverse set of viewpoints and expertise from across the DoD, academia, and commercialization sectors to explore these perspectives more in depth. To fully explore these perspectives, the attendees of the workshop shared their knowledge and expertise by participating in the following ways:

- defining appropriate notions and baseline attributes for operational trusted autonomy,
- identifying end-state goals,
- characterizing current state-of-the-art and gaps in knowledge,
- shaping requirements for future investments, and
- identifying effective strategies for transitioning ideas from the lab into practice through commercialization and workforce development." [3]

In addition to the technical autonomy knowledge, the IAA was specifically leveraged for their expertise in human-centered design methods and were charged as the planning and coordination lead. The IAA planning and coordination efforts needed to ensure that the right data was generated and collected throughout the summit to achieve outcome goals. Additionally, this meant that the IAA was also the lead in identifying the appropriate ideation activities and developing the flow of activities used to take the attendees throughout the summit. This included writing the facilitation plan and training a diverse set of collaborators to act as facilitators for each table of attendees during the ideation breakout sessions. The group of collaborators who facilitated the breakout sessions primarily came from the IAA, ICON, and NSIN. The objective of the breakout session was to be as generative as possible, across several different problem statements. From reviewing frequently deployed HCD tools, techniques and methodologies were considered that focused most on capturing the needs, desires, and meanings [17]. Ultimately, it was decided to predominantly leverage the brainstorming tool to facilitate a generative breakout session.

Brainstorming

Brainstorming is a divergent ideation method that can be leveraged by individuals or groups, where participants generate ideas around a central topic, following a set of ground rules to ensure a non-judgmental environment. While a common perception of brainstorming is that it is a simple activity with few to no rules, it is actually a complex social process requiring knowledge of social psychology, motivation, and culture [33].

The creativity within a group is influenced by their various cognitive, social, and motivational factors [34, 35]. These factors can range from negative social influences (apprehension, blocking), to positive social influence (accountability, competition), to positive cognitive

influences (attention, diversity, associations) [35]. The apprehension of sharing one's ideas, and thus making oneself vulnerable to critique, is an inhibitor of contribution and decreases individual motivation to contribute during creativity sessions [36]. Additionally, having to compete for speaking time in order to share their ideas (blocking) is also a motivation inhibitor [37, 38]. These inhibitory factors and influences can ultimately lead to group convergence [39], which is in direct contradiction to the goal of divergent generation. Conversely, there are positive influences that can drive positive group outcomes and generation. Increased accountability and a sense of (friendly) competition can increase both individual and group performance and outcomes [39]. Additionally, the cognitive diversity of participant knowledge and associations generated also contribute to both internal and external motivations, which help drive group performance [40].

It has been shown that the common verbal brainstorming method, where participants gather to share ideas face-to-face, is both generative and participant enjoyment and perceived effectiveness is high [41–43]. In groups where there is shared "talking time" in order to discuss ideas, there can be an issue of production blocking leading to decreased production [37, 38]. While waiting for their turn to speak, it is possible that participants may forget their original idea while listening to others share their idea, may instead fully focus on remembering others' ideas, and/or decide that their original idea is not worth sharing. Competing demands for participant attention and memory in listening to others while mentally holding on to their idea, can decrease production and make it difficult to build on the ideas of others [44].

A popular approach to counter this effect is to leverage the technique of brainwriting. Brainwriting is very similar to the verbal brainstorming method, but one key difference is that the participants write down their ideas on paper in some form [45–47]. By writing their ideas on paper, there is decreased risk of losing ideas while listening to and formulating responses to the ideas of others.

A crucial aspect to having positive generative brainstorming sessions is to create an atmosphere of acceptance of all ideas and to set ground rules of expected behavior during the session. The most popular rule is deferring judgment on the quality of ideas of others [33] but also of self. While there are other common rules that are used (go for quantity, build on the ideas of others, be visual), deferring judgment will have the largest impact on the success of the brainstorming session.

Additionally, brainstorming can be enhanced by leveraging a facilitator to enforce the rules and safe space, and also continue to drive discussion and focus on eliciting as many ideas as possible [33, 48]. To help additionally focus the participants, it's important to focus on one aspect of a problem or focus area at a time [34]. While not the full set of recommendations, the above-mentioned guidelines for brainstorming comprise the core of developing our approach.

Approach

Participants and Represented Sectors

A main driver of START was to convene and equitably solicit opinions from diverse stakeholders and experts from academia, DoD, and commercialization sectors. Mirroring the planning committee, the invited participants spanned across multiple technical (e.g., multi-disciplinary engineering domains) and non-technical (e.g., business development) backgrounds, working professions, areas of expertise, ethnicities, gender expression and identities, organizations, and sectors (government, academia, venture capitalists, and other entities). The summit hosted a total of 118 participants from 63 unique sector entities.

Due to the nature of the challenge space, all institutions represented at the event were based out of the United States. These participants included representatives from:

- 24 universities and academic institutions
- 16 U.S. Department of Defense (DoD) Units/Labs/Commands
- 10 venture capitalists
- 13 “other” entities which included organizations such as Federally Funded Research and Development Centers (FFRDCs), small businesses, etc.

A more thorough break down of participant breakdown by sector entities, and which specific entities engaged with and participated during the event, can be found in Appendix A.

Data Capture and Rules of Brainstorming

When brainstorming in a large and diverse group, all data (in the form of shared expertise or perspectives) was captured on sticky notes via a permanent marker. As mentioned previously, this method of data capture allows for rapid capture of thoughts, and flexibility to create and modify dynamic structures or clusters of information on the fly, as more data is added throughout the sessions. To make the brainstorming sessions as effective and open as possible for all attendees to share their perspective, the seven IDEO Rules of Brainstorming was leveraged for the breakouts, with an additional brainstorming rule included [49]. The brainstorming rules used were as follows:

- Defer Judgment: Creative spaces are judgment-free zones— they let ideas flow so people can build from each other’s great ideas.
- Encourage Wild Ideas: Embrace the most out-of-the-box notions. There’s often not a whole lot of difference between outrageous and brilliant.
- Build on the Ideas of Others: Try to use “and” instead of “but," it encourages positivity and inclusivity and leads to tons of ideas.
- Stay Focused on the Topic: Try to keep the discussion on target. Divergence is good, but you still need to keep your eyes on the prize.
- One Conversation at a Time: This can be difficult—especially with lots of creative people in a single room—but always think about the challenge topic and how to stay on track.
- Be Visual: Use colored markers and sticky notes. Stick your ideas on the wall so others can visualize them.

- Go for Quantity: Crank your ideas out quickly. For any 60- minute session, you should try to generate 100 ideas.
- (Extra, non-IDEO Rule) If It's Not on a Sticky Note, It Doesn't Exist: Sticky notes, not spoken words, are the main artifacts that are captured. Be sure your idea is on a sticky note for it to live past the conversation.

Theme Focus Areas and Summit Structure

Each of the three themes had a problem statement that directly provided the input and focus for each of the breakout brainstorming sessions, as seen in Table 1.

Table 1: Theme Problem Statements

Theme Focus Area	Problem Statement
Human-Centric	Formulate criteria and metrics to capture how humans perceive and trust autonomous systems, characterize methods to measure and modify human trust, and identify salient features of human decision making that will be relevant to design and operation of trusted autonomous systems.
Machine-Centric	Formulate criteria and metrics that capture trustworthiness of autonomous systems, characterize methods to measure trustworthiness of autonomous systems, and identify techniques for autonomous systems to dynamically improve trustworthiness.
Integration-Centric	Formulate criteria and metrics to evaluate the effectiveness of human-machine teams, characterize methods to measure bidirectional trust, and identify closed-loop techniques to dynamically adapt bidirectional trust for improving mission effectiveness.

For each of these theme focus areas, a similar process was followed for deep diving and exploring trust:

- Introduction Problem Statement: To start, each focus area was introduced with an introductory vision and problem statement specific to that individual perspective.
- Expert Panel Discussion: Immediately following, a panel discussion is held with leading experts to share their background, educate the attendees on their unique perspective, and inspire expansive thinking on each theme. Each of the three summit panels as part of the summit, focused on human, machine, and integration centric aspects of human-autonomy teaming, consisted of three panelists representing the Department of Defense, Academia, and the Commercialization Sector

- Breakout Session: Following each panel was a breakout session, where all participants contributed thoughts and viewpoints to answer key questions pertaining to each theme.

While the general flow is listed above, a more detailed description of the agenda and the components for each theme section is provided in Appendices B and C, respectively. Each section in Appendix C contains a description of the specific problem statement for each theme, how the problem statement was decomposed into three parts, the individuals who provided the vision or motivation for each section, as well as the accompanying panel members. Each breakout session focused specifically on one theme, and lasted 60 minutes in duration. Within the 60 minutes, all three breakout sessions followed the same general agenda, broken into 6 sections. All sessions had the same activities within it, with slightly varying time for each activity. The activities include an introduction to the activities, three rounds of brainstorming and sharing on different parts of the problem statement, followed by two rapid brainstorms with no sharing component (due to time limitations). Table 2 provides the timing breakdown for each 60-minute breaking session.

Table 2: Breakout Session Timing

Duration	Activity	Activity Description
5 minutes	Welcome and introduction	Welcome the participants at each table, do a quick round of introductions and set the scene for the hour activity
14 minutes	Heads Down Brainstorm 1 – Round 1	Using the prompt, the participants spend 2 minutes heads down (no talking), coming up with as many ideas as they can. One idea per sticky note.
	Share Out – Round 1	Each participant selects their top 1-2 ideas and provides a short (2-3 breath) share-out of their idea and why they think it's impactful. Their sticky note is moved to a large 3M sheet, and is either clustered with similar ideas or is placed in its own space. If time allows, a second round of sharing occurs.
	Thematic Clustering – Round 1	The facilitator finishes clustering the sticky note ideas into themes. Each theme should contain a sticky note header that describes the theme of the cluster in 1-4 words. Gather and keep the remaining sticky notes that were not shared out.
14 minutes	Round 2	Repeat the Heads Down Brainstorm, Share Out, and Thematic Clustering steps above for the new prompt.
14 minutes	Round 3	Repeat the Heads Down Brainstorm, Share Out, and Thematic Clustering steps above for the new prompt.
5 minutes	Rapid Brainstorm – Investment Area	Leveraging the three previous brainstorms as inspiration, the participants have ~2-3 minutes of heads down (no talking) focus to come up with as many ideas as possible for potential areas they would recommend investing

		resources in. Due to timing constraints, there is no share out. As they finish, the facilitator assembles and clusters the sticky notes onto the sheet.
8 minutes	Rapid Brainstorm - Near-Mid-Far barriers and enablers, potential path forward, and lessons learned	Leveraging the first three brainstorms as inspiration, the participants have ~2-3 minutes of heads down (no talking) focus to come up with as many ideas as possible for any or all of the following topics: for the near, medium, and long-term; what are the barriers vs. enablers, what path forward should we take, and what are some lessons learned (pitfalls). Due to timing constraints, there is no share out. As they finish, the facilitator assembles and clusters the sticky notes onto the sheet.

Digitally Capturing START Data

While there is immense benefit to capturing data manually during brainstorming, doing post-processing of analog data is much more cumbersome than using digital data. To that end, a member of each facilitation team digitized the contents of every sticky note that was generated for their group. This data was captured in an Excel sheet for each table, where each tab was a Brainstorm Topic-Theme Area intersection. For the 15 intersections of Brainstorm Topic-Theme Area across all of the groups, over 2000 thoughts and ideas were captured throughout the three generative brainstorming sessions.

The digitized data was transferred to an online collaborative whiteboarding tool (Miro) to allow for the consolidation and collocation of the large amounts of data collected. We recognize that while consolidating all data into a digital format allows for easier storage, limitations exist when leveraging traditional applications for handling large amounts of digital data (e.g., Microsoft Word, Excel, PowerPoint) due to the inability to freely manipulate and connect data points. However, we have found that using digital, online, collaborative tools, such as Miro, addresses/mitigates some of these shortcomings. Having such a powerful digital tool affords a space for distributed team members to directly interact and engage with the data, and provides a space for designers to collaborate together while they wrestle with the complexity of comprehending such large amounts of data. The goal at this stage of the design process is to deep dive into the data, identifying and uncovering connections and seemingly hidden meaning in the data, deriving clusters and themes from the disparate pieces of data through a process called synthesis.

Data Synthesis

Synthesis, which at times can be viewed as an unclear or ill-defined process, is a push towards the organization, reduction, and clarity of data or information in a way that reveals a sense of continuity and cohesion [50]. During this process designers steep themselves in the gathered data and work through the activity of organizing, categorizing, manipulating, parsing, and filtering data into a cohesive structure for information building [51]. Synthesis has thus been

defined as “an abductive sensemaking process of manipulating, organizing, pruning, and filtering data in the context of a design problem, in an effort to produce information and knowledge. [50]”

Sensemaking is a “a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively” [52]. It is an action-oriented process that people use to integrate experiences into their understanding of the world around [50], with a strong emphasis placed on finding patterns and relationships between whole or partial components of data. While there are variations and differences [53] in how sensemaking can be approached and achieved, across different methods there is less importance placed on being initially “accurate” and higher importance placed on giving tangible and abstract form to thoughts, reflections, ideas, and descriptions (on the patterns and relationships initially identified). After the externalization of data, the content is then organized and grouped, with each grouping given a label on the literal or implied relationship within that collection of data. These groups of data may be created, modified, abandoned, or any combination thereof several times to create an organized sense of data relationships. As all the generated content is related in some ways, it is not uncommon for important connections to be multifaceted, complex, and rooted in a common culture [50]. As it may be highly related to multiple groups, if needed, content may be duplicated for inclusion into the different emergent groupings.

With the sensemaking portion complete, the abductive reasoning step of synthesis can then take place. Abduction is the “step of adopting a hypothesis as suggested by the facts . . . a form of inference” [54]. Whereas deductive logic is the logic of “what should be” and inductive logic is the logic of “what is”, abductive logic allows for new knowledge or insights to be created from the facts presented, a “logic of what might be” [55].

It is worth clearly delineating the difference between deductive, inductive, and abductive logic:

- Deductive Logic (“what should be”): Deductive logic will guarantee the truth of the conclusion, based on the requirement that the premises presented are true. This type of logic is traditionally found in domains such as mathematics, and it is inherently self-contained and does not offer any new findings.
- Inductive Logic (“what is”): Inductive logic offers evidence that something might be true, with the basis being in structured experience. This type of logic is traditionally associated with scientific inquiry where subsequent experiences may invalidate or disprove initial hypotheses, but also not offer any new findings.
- Abductive logic (“what might be”): Abductive logic is the process of forming an explanatory hypothesis, which introduces new ideas or findings in the conclusion of the logical argument. It can be thought of as the “argument to the best explanation”, a “hypothesis that makes the most sense given observed phenomenon or data and based on prior experience” [50].

The process of design synthesis is a fundamental way to leverage the confines of a design problem to apply abductive logic [56]. While there are differences of opinion on if the abductive

suggestion is an act of a flash of insight [57], or merely appears to be a flash of insight but is rather a multi-step process that leads to an insight [58], both viewpoints agree that abductive reasoning is related to insight and creative problem solving. While there are different methods that can be used in different types of synthesis approaches [50], they commonly emphasize judging, prioritizing, and forming connections.

START Data Analysis and Synthesis Method

After digitizing all of the data, our team organized it by brainstorming theme area (human-centric, machine-centric, and integration-centric) on the virtual Miro whiteboard. Within each brainstorm theme area, "metric data" was reviewed and binned into "metric clusters" based on similarity in topic and commentary. Once all metric data was binned and tagged with a thematic cluster header, researchers looked across all metric cluster headers and further synthesized them into "metric themes." Figure 1 illustrates the organizing structure in which the metric data (gold notes) was synthesized into the metric clusters (green notes), which were then further consolidated into metric themes (black note).

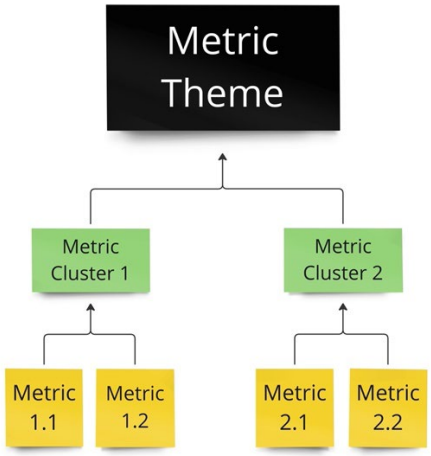


Figure 1: Metric Data to Metric Theme Mapping Diagram

Next, the corresponding methods (of which the data for was derived during subsequent brainstorms during the START event) linked with each metric theme and associated metric clusters were identified and aggregated. A generic example of this data organization can be seen in Figure 2, where the corresponding "method data" is noted in shades of blue and teal.

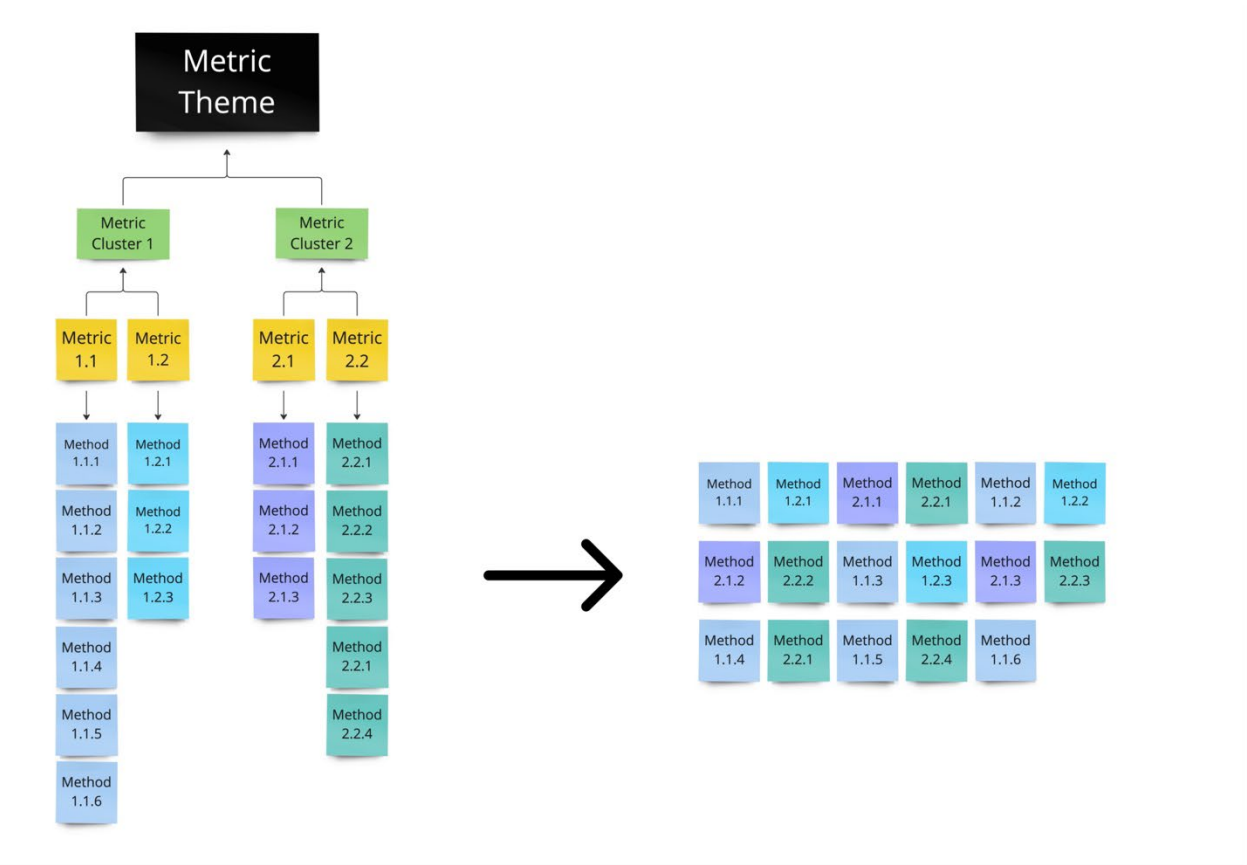


Figure 2: Metric Data to Method Data Diagram

Similar to the metric data analysis, the method data was reviewed and binned into “method clusters” based on similarity in topic and commentary, as seen in Figure 3 .

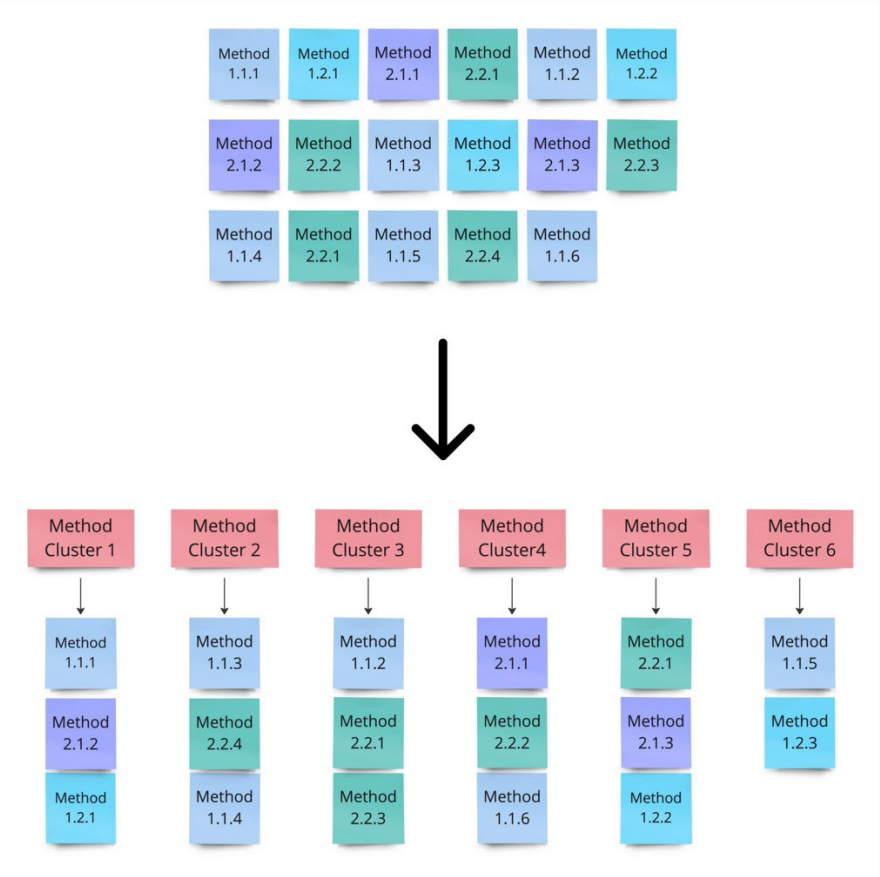


Figure 3: Method Data to Method Cluster Diagram

Once all method data was binned and tagged with a method cluster header, researchers looked across method cluster headers within each brainstorm theme area in order to identify points of overlap, connection, and potential gaps. The data consolidation and connection points resulted in “abstracted themes,” as illustrated in Figure 4.

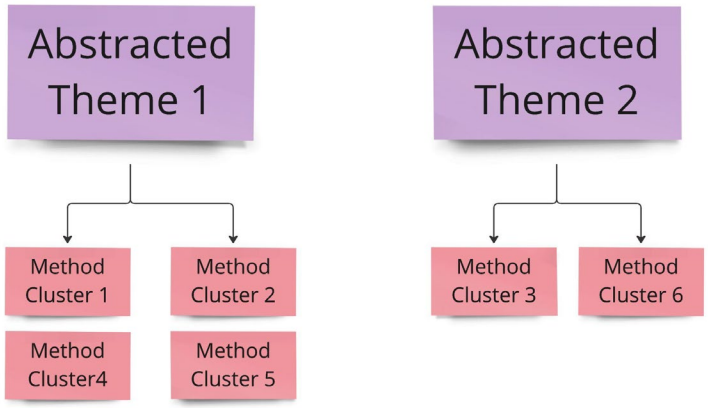


Figure 4: Method Clusters to Abstracted Themes Diagram

These metric, method, and abstracted themes were the foundational concepts that researchers further synthesized and leveraged to derive insights related to specific quantitative and qualitative metrics for measuring trust impact factors and design guidelines for systems and human-machine interactions.

Results and Discussion

A variety of different themes were derived from the START data and were categorized as either "cross-cutting themes" (e.g., themes that were seen repeatedly across the 3 different brainstorm focus areas) or "unique themes" (e.g., themes that did not cut across the 3 different brainstorm focus areas and, instead, were unique to 1 specific brainstorm focus area). Cross-cutting themes included concepts such as:

- Human Intervention and Oversight
- Human-Machine Feedback Loops
- Ability for System to Match Expected Behaviors
- Machine Explainability and Intention Transparency
- Human-Machine Team Performance

These themes highlight suggested influences on operational trust in mission autonomy and therefore opportunity areas for further investigation and/or development of specific metrics.

To identify these specific metrics, the data which support the cross-cutting themes was further analyzed, using the process described above. These specific metrics, along with explanation descriptions, can be found in the section below.

In addition, because the dataset was so large and complex, the synthesis process also produced a set of design and implementation guidelines, which outline some key considerations for the future design and development of mission autonomy. These guidelines are also presented in the subsequent section below.

Metrics

The data linked to the cross-cutting metric themes outlined above was further analyzed and synthesized into specific metrics for better understanding how factors which contribute to operational trust could be measured (either quantitatively or qualitatively) through a variety of techniques and experiments. These metrics (in bold), with brief descriptions, can be found below.

Human Intervention and Oversight

- **Number of Times System "Calls for Help":** How often the machine prompts a human operator and requires a human operator's attention.
- **Number of Times A Human Provides Guidance:** How often a human operator is required to provide direction for the machine (and how long the machine can act without human operator provided direction).

- **Number of Times a Human Declines System Recommendations/Decisions:** How often the machine makes decisions that must be overwritten by a human operator.
- **Number of Times A Human is Required to Intervene/Takeover:** How often a human operator is required to take direct control of the machine's actions.

Human-Machine Feedback Loop

- **Ability to Query:** How easily a human operator can prompt the machine for clarification on the machine's understanding or intentions and receive an understandable response.
- **Constant Feedback and Communication Loop:** In what manner and how often the machine notifies a human operator of its situational understanding of the environment, operational planning, or detected error.
- **Communication Prior to Action:** How much and how often the machine informs a human operator of its understanding and intentions before performing a task.
- **Information Sharing Optimization**: How well the machine can optimize the way it shares information based on its current human operator.

Ability for Machine to Match Expected Behaviors

- **User Expectations vs. Machine Actions:** How well the functionality and performance of the machine matches the expectations of the human operator and meets the human operator's needs.
- **Design Expectations vs. Machine Actions:** How closely the functionality and performance of the machine matches the expectations of how it was designed to complete tasks.

Machine Explainability and Intention Transparency

- **Auditability:** How robustly the machine collects data on its detection of the operating environment and its decision-making methodology and how well it can communicate that understanding to the human end user, when/if prompted.
- **Understanding**: How well the machine can gather knowledge of the human users through both direct communication (intentionally sent messages) and indirect communication (observed states, traits, or behaviors).
- **Transparency:** How visible the internal understanding and decision-making methodology of the machine is to both internal team members (either humans or machines) and outside observers.

Human-Machine Team Performance

- **Accuracy**: How well the human-machine team can correctly perform its task(s).
- **Consistency & Repeatability:** How well the human-machine team can perform its task(s) multiple times in the same manner.
- **Safety:** How well the human-machine team can prevent itself and its components from being harmed while performing its tasks.

Although we present a variety of useful metrics, which could be leveraged to measure and better understand aspects of operational trust, it is important to note that these do not represent a complete set and are not the only metrics that are useful or contribute to better understanding technical systems and their interactions with the world. These metrics were synthesized based on the data from START, and therefore, this set acts as a representation of the voices and perspectives that participated in the event.

In addition, it is important to consider that although a set of metrics are presented, trust is a very nuanced topic and it cannot simply be "measured." Trust is an emotion that is based on the complexities of human experience and individuality, and we should not attempt to investigate nor measure it through a myopic, singular lens [59]. Rather, operational trust must be thought of as a fluid construct that is considered and designed for from the very beginning of a system's life cycle.

Design and Implementation Guidelines

In addition to identifying specific quantifiable metrics, the START data was further synthesized and molded into a set of design and implementation guidelines. These guidelines aim to act as starting points and initial considerations for designing, developing, and implementing *trusted* mission autonomy. These guidelines were organized into four core perspectives: the human, the technical system, the human-machine team, and the ecosystem.

While there is a clear distinction between the human and the technology of a human-machine team, the boundary that separates the "machine teammate" from the rest of the supporting technology infrastructure is much less clear. We recognize that a variety of definitions and organizing structures exist. Therefore, for the purposes of this paper, we defined the four perspectives as:

- **Human:** Human refers to the human members of a human-machine team.
- **Technical System:** The technical system refers to the "machine teammate" or intelligent autonomous system. This system can include the machine hardware that the human interacts with along with the logic and reasoning algorithms/software that run the hardware.
- **Human-Machine Teaming:** Human-machine teaming refers to the shared communication, information, and control between the human and the technical system. This teaming includes interactions between the human and technical system along with instances where they share information with one another or change their behavior according to their understanding of one another.
- **Ecosystem:** Ecosystem refers to everything that supports the human-machine team which is not a human nor the technical system. This perspective includes but is not limited to both the physical and digital infrastructure, stewards, established ethics and policy, lifecycle support, maintenance, and training associated with the human-machine team along with any performance analysis and assessment performed for the human-machine team.

For the purposes of our finding, the perspectives are thought about in the organizing construct seen in Figure 5.

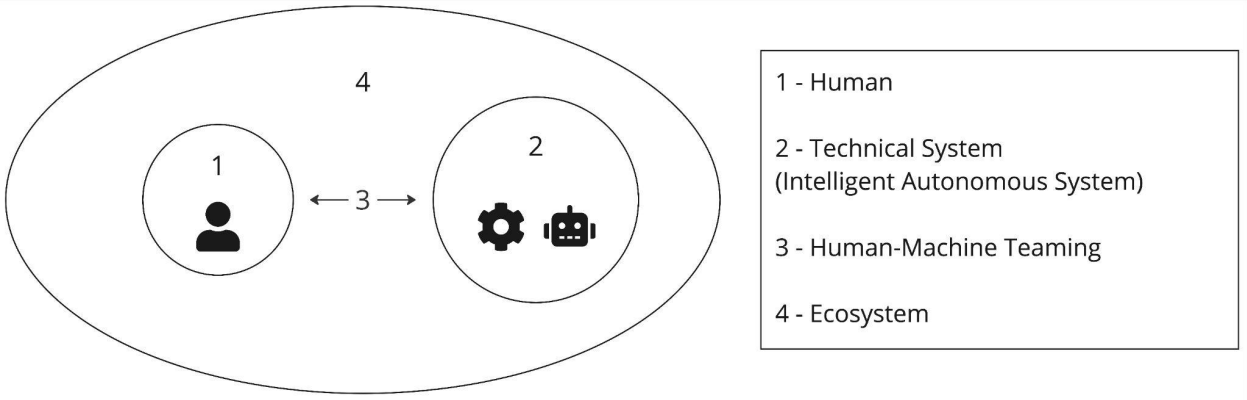


Figure 5: Perspective Organizing Structure

Similar to the set of metrics presented above, it is important to note that these guidelines were derived from the same START dataset as the metrics, and therefore, do not represent a complete nor comprehensive set of guidelines. In addition, these guidelines are not presented as “needs'' or “requirements” but rather presented as “should” or “could” statements, to aid in prompting thinking and starting conversations around these nuanced and complex topics of operational trust and mission autonomy.

Some of these guidelines and themes are more strongly linked to a certain perspective (e.g., topics around trust are typically linked to the “human” perspective, since it is an emotion based on human experience and perspective). However, there are also themes that are seen as common threads across multiple perspectives (e.g., highlighting the benefit of adaptability spans across all four perspectives). Both of these linkages are interesting - the more unique linkages showcase the need for specific focus on a certain topic/perspective (such as leveraging HCD methods) whereas themes that touch multiple different perspectives highlight the need to approach these topics with a more holistic mindset. The themes that cut across multiple perspectives also demonstrate how a single design principle might have an impact across a variety of attributes.

Below, a high-level summary of these guidelines can be found under each thematic section. A complete list of the guidelines can be found in Appendix D.

Safety and Testing

Humans should have the ability to maintain some level of control over the system - with the ability to intervene and override system actions, when and if appropriate. The human-machine team should be able to detect if and when an error occurs and be able to react. In turn, these technical systems must be developed with resilience mechanisms in mind - with the ability to operate in its pre-defined operational design domain yet fail gracefully, if necessary. In addition, to ensure safety and robustness, systems should be extensively tested and trained in a variety of both simulated and real-world environments and testbeds.

Data Collection and Analysis

Technical systems should have on-board sensing capabilities to determine operator state (e.g., stress level, fatigue) and the ability to analyze and validate collected data in-situ. The surrounding ecosystem that supports the human-machine team should have appropriate digital infrastructure (e.g., data pipelines and storage) to enable collection and measurement of information about the environment.

Performance Measurements and Comparisons

It is important for humans to have a clear understanding of a technical system's capabilities, such as the ability for it to complete certain tasks and what conditions and environments it can operate optimally in. The technical system should also be able to determine and communicate to the human which tasks are completed "successfully." Predefined standards of performance, specific for given tasks and environments, should be determined for a given system, and the ecosystem should have procedures in place for assessing technical system and/or human-machine team performance compared to these baselines.

Human-Machine Interactions

To build trust between humans and the system and enhance human-machine teaming, participants indicated that the ability for humans to observe other humans interacting with the system might benefit the overall interaction. In addition, there should be shared and agreed upon expectations for how a system will act and how it will communicate those actions, or deviations from expected actions, to the human. To help facilitate this shared understanding, humans should also be a part of the system design and development through co-design and feedback mechanisms. The technical system, having been designed with the human operator involved in the process, should have the capability to act as a team member and adapt to the informational and control needs of human teammates that it identifies through observation and communication.

Transparency and Explainability

In addition to the humans having insight and understanding of reasoning and intention behind the technical system's actions, they should also have an ability to identify if/when a technical system is being tampered with (e.g., spoofing). Achieving this understanding could be possible through repetitive and prolonged interactions between the human and machine, however, also providing ample documentation and training to humans interacting with the technology could be beneficial for spotting unusual or deviations from normal behavior.

System Confidence and Reliability

Supporting ecosystems should support the measurement and evaluation of system performance over time. These calculations can include such metrics as system error, accuracy, safety, reliability, and confidence. These calculated metrics should be compared and evaluated based on preestablished taxonomies and criteria for acceptable/optimal system performance.

Human Trust and Reliability

As mentioned in other thematic categories, participants highlighted that humans should be kept in the loop and fully informed of technical system understanding of the environment and explanation of actions in order to foster trust. In the near-term, systems are not completely independent nor inherently trusted, therefore, humans must be kept in the technology loop in case they must take over system control. During this time, the ecosystem should support the ability to capture and account for indications of human trust over time. These measurements can help inform design and implementation decisions.

User Training and Adoption

Users should have a baseline familiarity with the system, task at hand, and the system's ability to complete the task. This familiarity could be a product of system use over time (e.g., repetitive exposure to the system) or through training programs dedicated to a particular system or technology attribute. As humans gain more exposure to and experience with a system, better understanding its appropriate use and limitations, their confidence in the system and willingness to leverage similar systems should also increase. Note, training is not a replacement for real-world, hands-on experience with the system, and it should not be expected that training will be the only means for increased system familiarity and confidence.

Bias Detection/Mitigation

Identifying and mitigating bias from technical systems is crucial for system effectiveness and user trust in the system. Therefore, technical systems should have the ability to identify and detect biases in both data and human mental models. In addition, the ecosystem should be able to help support bias identification and mitigation in both the humans and machines through mechanisms such as bias mitigation, monitors, governance, and policy.

Adaptability and Flexibility

Because the world is such an ever-changing and complex place, it is important to design for adaptability and flexibility. Users should have the ability to inform modification and customization of technical system features based on their preferences. In turn, the technical system should enable the customization while also adapting its features based on task complexity, experience level of the human (novice vs. expert user, familiarity with the system, etc.), and trust level. Standards and regulations should lay out firm guidelines for system development, yet also be flexible to enable this adaptability.

Conclusion

In summary, the process outlined above highlights how HCD methodologies can be applied to large and ambiguous design challenges. In this specific case, the HCD approach was applied to the research question: "how might we identify and/or develop metrics of operational trust in mission autonomy?"

The HCD approach worked well for our particular challenge because it enabled us to explore trusted mission autonomy at a more holistic scale and provided the process framework for capturing ideas and perspectives of a diverse stakeholder group. Leveraging the power of brainstorming as the main ideation activity allowed us to be highly generative in this divergent phase of the project. Thoughtful consideration was put into setting up the appropriate structure, guidelines, and practices for the generative sessions to great success.

In addition to outlining the method, this paper also presents findings from the data collected at START. These findings illustrate how across multiple social, cultural, and technical contexts, ideating with a diverse group of people can provide a variety of insights, in both breadth and depth. The findings and insights of the START dataset provided both specific metrics that can be leveraged to evaluate system performance and interactions with human users, as well as overarching design and implementation recommendations that can inform technical system design, development, and implementation. These guiding principles and frameworks aim to act as guides and conversation starters, as we continue to navigate the nuanced topic of operational trust in mission autonomy.

We recognize that HCD is not the right solution to every problem. However, when applied to the right challenge, it can help create products, services, and systems that are not only functional and efficient, but also meaningful and impactful across a broader range of people within the entire system environment.

Acknowledgements

We would like to acknowledge Dr. Kim Sablon and OUSD RDT&E for funding this effort.

We also like to acknowledge Dr. Jaret Riddick, author of the OPTIMA initiative and the driving force behind START. Thank you to the START organizing committee for their thoughtful collaboration and critical contributions: from Purdue University we would like to thank Prof. Shreyas Sundaram (Electrical and Computer Engineering & START Co-Chair), Prof. Shaoshuai Mou (Aeronautics and Astronautics & START Co-Chair), Prof. Hubo Cai (Civil Engineering), Prof. Neera Jain (Mechanical Engineering), Prof. Mahsa Ghasemi (Electrical and Computer Engineering), Dr. Abhijit Karve (Director of Business Development, Office of Technology Commercialization), Prof. Brandon Pitts (Industrial Engineering), and Prof. Richard Voyles (Purdue Polytechnic). From NSIN we would like to thank Kedar Pavgi. From the Office of the Undersecretary of Defense we would like to thank Michael DiPaolo, Shannon Arnold, and Dana Franz. Special thanks to Mary Ann Bobillo, Heather Anthrop, and Reed Skaggs, who provided event and logistics support for START.

We would also like to acknowledge the START moderators and panelists, a full list of which can be found in Appendix C.

References

[1] Poornima Madhavan and Douglas A. Wiegmann. 2007. Similarities and differences between human-human and human-automation trust: An integrative review. Theoretical Issues in Ergonomics Science 8, 4 (May 2007), 277–301. https://doi.org/10.1080/14639220500337708.
[2] Matthew Johnson and Alonso Vera. 2019. No AI Is an Island: The Case for Teaming Intelligence. AI Magazine 40, 1 (March 2019), 16–28. https://ojs.aaai.org/aimagazine/index.php/aimagazine/issue/view/225.
[3] Jaret Riddick. 2022. Operational Trust in Mission Autonomy (OPTIMA). National Security Innovation Network. https://www.nsin.mil/assets/downloads/Operational_Trust_in_Mission_Autonomy_START_20220628.pdf
[4] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Not So Different After All: A Cross-Discipline View of Trust. Academy of Management Review 23, 3 (July 1998), 393–404. https://journals.aom.org/doi/10.5465/amr.1998.926617.
[5] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies 27, 5-6 (November 1987), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5
[6] Nancy K. Lankton, D. Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust:
Rethinking trust in technology. Journal of the Association for Information Systems 16, 10 (October 2015), 1–33. https://aisel.aisnet.org/jais/vol16/iss10/1
[7] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society 5, 1 (March 2018). https://doi.org/10.1177/2053951718756684
[8] Kristin E. Schaefer, Jessie Y.C. Chen, and P.A. Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human Factors 58, 3 (March 2016), 377–400. https://doi.org/10.1177/0018720816634228
[9] John D. Lee, and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. Human Factors 46, 1 (August 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
[10] Joseph B. Lyons, Kevin T. Wynne, Sean Mahoney, and Mark A. Roebke. 2019. Chapter 6 - Trust and Human-Machine Teaming: A Qualitative Study. Academic Press, 101–116. https://doi.org/10.1016/B978-0-12-817636-8.00006-5
[11] Chad C. Tossell, Boyoung Kim, Bianca Donadio, Ewart J. de Visser, Ryan Holec, and Elizabeth Phillips. 2020. Appropriately Representing Military Tasks for Human-Machine Teaming Research. In Proceedings of the HCI International 2020 - Late Breaking Papers: Virtual and Augmented Reality (Stephanidis, C., Chen, J.Y.C., and Fragomeni, G. eds.). Lecture Notes in Computer Science, vol 12428. Springer, Cham, Article 19, 1–7. https://doi.org/10.1007/978-3-030-59990-4_19
[12] Nicole Blatt. 2004. Operational Trust: A New Look at the Human Requirement in Network Centric Warfare. In Proceedings of the 9th International Command and Control Research and

Technology Symposium Coalition Transformation: An Evolution of People, Processes, and Technology to Enhance Interoperability. https://apps.dtic.mil/sti/pdfs/ADA466612.pdf
[13] Gari Palmer, Anne Selwyn, and Dan Zwillinger. 2014. The “Trust V” - Building and measuring trust in autonomous systems. Raytheon Co., Tewksbury, MA, Tech. Rep., 2014. https://link.springer.com/chapter/10.1007/978-1-4899-7668-0_4
[14] The Johns Hopkins Institute for Assured Autonomy. 2022. Johns Hopkins IAA - About. Accessed: 2/16/2022. https://iaa.jhu.edu/about/
[15] Jon Kolko. 2012. Wicked Problems: Problems Worth Solving. Stanford Social Innovation Review 10, 3 (March 2012), 1–5. https://doi.org/10.48558/1REG-NX98
[16] Susan Gasson. 2003. Human-centered vs. user-centred approaches to information system design. The Journal of Information Technology Theory and Application (JITTA) 5, 2 (January 2003), 29–46. https://cci.drexel.edu/faculty/sgasson/papers/SG-JITTA.pdf
[17] Joseph Giacomin. 2014. What Is Human Centered Design? The Design Journal 17, 4 (April 2014), 606–623. https://doi.org/10.2752/175630614X14056185480186
[18] Lucy Suchman. 2007. Human-Machine Reconfigurations: Plans and Situated Actions, 2nd ed. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511808418
[19] Marc Steen. 2011. Tensions in human-centred design. CoDesign 7, 1 (June 2011), 45–60. https://doi.org/10.1080/15710882.2011.563314
[20] Patrick W. Jordan. 2002. Human factors for pleasure seekers. In Design and the Social Sciences: Making Connections (Frascara, J. ed.). Taylor and Francis, London, UK, 9–23.
[21] Ezio Manzini. 2014. Making things happen: Social innovation and design. Design issues 30, 1 (May 2014), 57–66. https://www.jstor.org/stable/24267025
[22] Richard Buchanan. 1992. Wicked Problems in Design Thinking. Design Issues 8, 2 (January 1992), 5–21. https://doi.org/10.2307/1511637
[23] John C. Camillus. 2008. Strategy as a wicked problem. Harvard Business Review 86, 5 (May 2008), 98–109. https://hbr.org/2008/05/strategy-as-a-wicked-problem
[24] Horst Rittel and Melvin Webber. 1973. Planning problems are wicked. Polity 4, 155–169. http://www.jstor.org/stable/4531523?origin=JSTOR-pdf
[25] Jeanne Liedtka. 2015. Perspective: linking design thinking with innovation outcomes through cognitive bias reduction. Journal of Product Innovation Management 32, 6 (November 2015), 925–938. https://doi.org/10.1111/jpim.12163
[26] Fabrizio Ceschin and Idil Gaziulusoy. 2016. Evolution of design for sustainability: From product design to design for system innovations and transitions. Design studies 47 (November 2016), 118–163. https://doi.org/10.1016/j.destud.2016.09.002
[27] Luca Rota, Yanjun Zohu, and Svenja Paege. 2019. Sustainable Product-Service System Design from a strategic sustainable development perspective. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1340289&dswid=-8792
[28] Frederick A. Muckler and Sally A. Seven. 1992. Selecting Performance Measures: “Objective” versus “Subjective” Measurement. Human Factors 34, 4 (November 1992), 441–455. DOI:https://doi.org/10.1177/001872089203400406
[29] SimpleKPI. 2021. What are Metrics and Measures - Explanation and Examples. Accessed: 2/16/2022. Available: https://www.simplekpi.com/Blog/metrics-and-meassures-a-definitive-guide

[30] The Johns Hopkins Institute for Assured Autonomy. 2021. Johns Hopkins IAA - Video: IAA Human-Centered Design Team At AUVSI XPonential 2021. Accessed: 2/16/2022. Available: https://iaa.jhu.edu/video-iaa-human-centered-design-team-at-auvsi-xponential-2021/
[31] Purdue University. 2022. Institute for Control, ICON Optimization and Networks - About. Accessed: 2/22/2022. Available: https://engineering.purdue.edu/ICON
[32] National Security Innovation Network. 2022. Accessed: 2/22/2022. Available: https://www.nsin.mil/
[33] Chauncey Wilson. 2013. Brainstorming and beyond: a user-centered design method. Newnes. https://doi.org/10.1016/C2012-0-03533-8
[34] Paul B. Paulus and Vincent R. Brown. 2003. Enhancing ideational creativity in groups. Group creativity: Innovation through collaboration, 110–136. https://doi.org/10.1093/acprof:oso/9780195147308.003.0006
[35] Paul B. Paulus, Vicky L. Putman, Karen Leggett Dugosh, Mary T. Dzindolet, and Hamit Coskun. 2002. Social and cognitive influences in group brainstorming: Predicting production gains and losses. European review of social psychology 12, 1, 299–325. https://doi.org/10.1080/14792772143000094
[36] Rolf Faste. 1993. An Improved Model for Understanding Creativity and Convention, in Cary A. Fisher (ed.), ASME Resource Guide to Innovation in Engineering Design, American Society of Mechanical Engineers. http://fastefoundation.org/publications/an_improved_model.pdf
[37] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. Journal of personality and social psychology 53, 3 (September 1987), 497–509. https://doi.org/10.1037/0022-3514.53.3.497
[38] Michael Diehl and Wolfgang Stroebe. 1991. Productivity loss in idea-generating groups: Tracking down the blocking effect. Journal of Personality and Social Psychology 61, 3 (September 1991), 392. https://psycnet.apa.org/doi/10.1037/0022-3514.61.3.392
[39] Paul B. Paulus and Jared B. Kenworthy. 2019. Effective brainstorming. In The Oxford Handbook of Group Creativity and Innovation, 287-386. https://psycnet.apa.org/doi/10.1093/oxfordhb/9780190648077.013.17
[40] Carsten De Dreu, Bernard Nijstad, and Daan van Knippenberg. 2008. Motivated information processing in group judgment and decision making. Personality and Social Psychology Review 12, 1 (February 2008), 22-49. https://doi.org/10.1177/1088868307304092
[41] Paul B. Paulus and Mary T. Dzindolet. 1993. Social influence processes in group Brainstorming. Journal of Personality and Social Psychology 64, 4, 575. https://psycnet.apa.org/doi/10.1037/0022-3514.64.4.575
[42] Paul B. Paulus, Timothy S. Larey, and Anita H. Ortega. 1995. Performance and perceptions of brainstormers in an organizational setting. Basic and Applied Social Psychology 17, 1-2, 249-265. https://psycnet.apa.org/doi/10.1207/s15324834basp1701&2_15
[43] Wolfgang Stroebe, Michael Diehl, and Georgios Abakoumkin. 1992. The illusion of group effectivity. Personality and Social Psychology Bulletin 18, 5 (October 1992), 643-650. https://doi.org/10.1177/0146167292185015
[44] Bernard A. Nijstad, Wolfgang Stroebe, and Hein Lodewijkx. 2003. Production blocking and idea generation: Does blocking interfere with cognitive processes? Journal of Experimental Social Psychology 39, 6, 531-548. https://doi.org/10.1016/S0022-1031(03)00040-4

[45] Horst Geschka. 1993. The development and assessment of creative thinking techniques: A German perspective. In Nurturing and Developing Creativity: The Emergence of a Discipline, 215-236. https://doi.org/10.1111/j.1467-8691.1996.tb00125.x
[46] G. Goodman. 1995. Brainwriting: What to do when there's not a cloud in the brainstorming sky. In Marketing Encyclopedia: Issues and Trends Shaping the Future, 40-46.
[47] Arthur B. VanGundy. 1981. Techniques of Structured Problem Solving. New York: Van Nostrand Reinhold. https://link.springer.com/book/9780442288471
[48] Robert I. Sutton and Andrew Hargadon. 1996. Brainstorming Groups in Context: Effectiveness in a Product Design Firm. Administrative Science Quarterly 41, 4 (December 1996), 685-718. https://psycnet.apa.org/doi/10.2307/2393872
[49] IDEO U. 2017. 7 Simple Rules of Brainstorming. Accessed: 2/22/2022. Available: https://www.ideou.com/blogs/inspiration/7-simple-rules-of-brainstorming
[50] Jon Kolko. 2010. Abductive Thinking and Sensemaking: The Drivers of Design Synthesis. Design Issues 26, 1 (January 2010), 15-28. https://doi.org/10.1162/desi.2010.26.1.15
[51] Jon Kolko. 2007. Information Architecture and Design Strategy: The Importance of Synthesis during the Process of Design. In IDSA 2007 Educational Conference Proceedings (San Francisco: IDSA). https://www.jonkolko.com/writing/information-architecture-and-design-strategy
[52] Gary Klein, B. Moon, and Robert Hoffman. 2006. Making Sense of Sensemaking 1: Alternative Perspectives. Intelligent Systems (IEEE) 21, 4 (July/August 2006), 71. https://doi.org/10.1109/MIS.2006.75
[53] Brenda Dervin. 2003. Sense-Making's Journey from Metatheory to Methodology to Methods: An Example Using Information Seeking and Use as Research Focus. In Sense-Making Methodology Reader (Cresskill, NJ: Hampton Press), 141–146.
[54] Charles S. Peirce. 1998. On the Logic of Drawing History from Ancient Documents. In The Essential Peirce: Selected Philosophical Writings, 1893–1913, edited by Peirce Edition Project (Bloomington: Indiana University Press), 95
[55] David Dunne and Roger Martin. 2006. Design thinking and how it will change management education: An interview and discussion. Academy of Management Learning & Education 5, 4 (December 2006), 512–523. https://psycnet.apa.org/doi/10.5465/AMLE.2006.23473212
[56] Richard Coyne. 1988. Logic models of design. Pitman.
[57] Charles S. Peirce. 1988. Pragmatism as the Logic of Abduction. In The Essential Peirce: Selected Philosophical Writings, 1893–1913, edited by Peirce Edition Project (Bloomington: Indiana University Press), 227
[58] Philip N. Johnson-Laird. 2005. The Shape of Problems. In The Shape of Reason: Essays in Honour of Paolo Legrenzi, edited by V Girotto (Psychology Press), 3–26.
[59] Frederik Schwerter & Florian Zimmermann. 2020. Determinants of trust: The role of personal experiences. Games and Economic Behavior, 122, 413-425. https://doi.org/10.1016/j.geb.2020.05.002

Appendix

A. Participants by Sector Demographics

The 118 diverse participants came from 63 unique sector entities, and the demographics of participant by sector and entity are described below.

There was a total of sixty-eight participants from twenty-four academic institutions, which included: Arizona State University, Carnegie Mellon University, Clemson University, Cornell University, Drexel University, Duke University, George Mason University, Georgia Institute of Technology, Harvard University, Johns Hopkins University, Northeastern University, Ohio State University, Purdue University, University of California Santa Cruz, University of Central Florida, University of Florida, University of Kansas, University of Illinois at Urbana-Champaign, University of Louisville, University of Massachusetts Lowell, University of Michigan, University of Wisconsin – Madison, University of Virginia, Vanderbilt University.

There was a total of twenty-three participants from sixteen Department of Defense Units/Labs/Commands, which included: AFRL, ARL, AFOSR, CDAO, DARPA, Defense Innovation Unit, DEVCOM ARL, GVSC, NAVSEA, NSIN, OSD, OUSD(R&E), TRMC, USAF, US Navy
There was a total of twelve participants from ten ventures, which included: AIN Ventures, America's Frontier Fund, Beacon Global Strategies, Decisive Point, Dioltas, gener8tor, IEDC, Pine Grove Holdings, Squadra.

There was a total of fifteen participants from thirteen “other” entities (FFRDC, Small business, etc.), which included: ARCUAS, Boeing, Codex Labs LLC., Edge Case Research, HRL, Lewis-Burke, M-Vision Inc., Mile Two LLC., MITRE, NASA, Nusura, Ubicept.

B. Detailed Facilitator Guide

Below is a detailed walk-through description for all of the activities mentioned above. Included in this walkthrough is the objective for each activity, as well as logistics or instructions for the facilitators to use and follow as needed. The prompts used are described in the detailed section for each theme in follow on sections.

Welcome, introduction to table and activity [5 minutes total]

- Facilitation Logistics
    - Greetings and welcome to the table
    - Individual 1-breath introductions – with one normal breath, each person gives a very quick introduction on themselves (name, org, job, etc.)
    - Introduce goal of collaborative session

First round brainstorm [14 minutes total]

- Facilitation Logistics
    - (2 min) Brainstorm Heads Down: Using the prompt, instruct the participants to spend the next 2 minutes heads down (no talking), and coming up with as many ideas as they can. One idea per sticky note.

- (10 min) Round Robin Top Share-out: Tell the participants to look at their ideas, and to select the top 1-2 ideas. Go around the table and have each person hand the facilitator their sticky note to put on the large 3M sheet, while providing a short (2-3 breath) share of their idea and why they think it's impactful.
    - It's common for people to want to share out a lot of details – hence the 2-3 breath maximum for sharing out the why.
    - As each person is sharing out their idea, remind them that we will capture and keep all of the sticky notes, but first round or two is to ensure the highest priority ones are collected and noted.
    - As the sticky notes are being shared, the facilitator should cluster the sticky notes into common themes. This will happen throughout the share process.
    - After the first round of ideas have been shared, if there is time, either do another round around the table, or open it up for quick discussion. What things/ideas jumped out as interesting or surprising?
- (2 min) Finish Clustering: If needed, the facilitator (with help from a participant if desired) should finish clustering the sticky notes into themes.
    - Create a header sticky note that in 1-4 words describes the theme of the cluster. Draw a border box around this sticky note to differentiate the sticky note as a header.
    - Gather and put on the sheet (if there is space) the remaining sticky notes that were not shared out. If there is no space, put them into a pile and set them to the side.

Second and third round brainstorms [14 minutes each]

- Repeat Round 1 procedure for Brainstorming Rounds 2 & 3

Fourth round – lightning brainstorms [13 minutes total]

- Activity Objective: This final round consists of two lightning brainstorms. As the focus of time is spent on the previous three brainstorms, these rounds will be very condensed and will not have a share out component. The goal is again to be very generative and capture large amounts of information in a very short time.
    - Prompt/Topic 1: What investments should be made?
    - Prompt/Topic 2: For the near, medium, and long-term; what are the barriers vs. enablers; what path forward should we take, and what are some lessons learned (pitfalls)?
    - Facilitation Logistics:
        - (5 min) Rapid Brainstorm 1: Introduce Prompt/Topic 1 from above. Give the participants ~2-3 minutes of heads down (no talking) focus, to come up with as many ideas as possible. There will be no share out. As they finish, hand in the sticky notes to the facilitator, who will assemble and cluster the sticky notes onto the sheet.
        - (8 min) Rapid Brainstorm 2: Introduce Prompt/Topic 2 from above. Give the participants ~2-3 minutes of heads down (no talking) focus, to come

up with as many ideas as possible. There will be no share out. As they finish, hand in the sticky notes to the facilitator, who will assemble and cluster the sticky notes onto the sheet. This brainstorm has more components so there is a bit more time allocated here.

C. Detailed Theme Description

Theme 1 – Human-Centric

For this theme, the problem statement used to drive the session was "Formulate criteria and metrics to capture how humans perceive and trust autonomous systems, characterize methods to measure and modify human trust, and identify salient features of human decision making that will be relevant to design and operation of trusted autonomous systems."

Mr. Robert Walker from the Naval Surface Warfare Center (NSWC), Crane Division provided the vision and motivation for the human-centric aspect of trusted autonomy through a roughly 10-minute presentation of remarks. At the time of the summit, Mr. Walker served as the newly elected Chief Technology Officer at NSWC, responsible for internal science and technology investments and innovation pilot programs, amongst other duties. Mr. Walker led projects for the military and other government agencies and has helped launch programs, technology, and partnerships internally and externally that are critical to national security. Additional information about Mr. Walker can be found at: https://www.navsea.navy.mil/Media/News/Article/1675685/nswc-crane-selects-new-chief-technology-officer/

For the human-centric theme, the panel to drive the education and inspiration on this topic consisted of the following members:

- Moderator: Prof. Brandon Pitts, Purdue
- DoD Panelist: Dr. Joe Lyons, AFRL
- Academic Panelist: Prof. Robert Proctor, Purdue
- VC Panelist: Mr. Tommy Hendrix, Decisive Point

The questions posed to the panelist to answer, and expound upon, are listed below.

- How do you define human trust in the context of human-automation/autonomy teaming?
- What makes it difficult to define (trust) metrics and what are the mature ways to measure/measurement trust?
- What are the main features/aspects of human trust that need to be understood or measured to design better automation?
- Trust has been studied since the 90's. In your opinion, …
- What progress/successes have we made
- What top challenges and/or barriers still remain
- Towards measuring and operationalizing trust in real-world operations?

- What advice do you have for how academia, DoD, and VC can work together to further the field of human-automation/autonomy teaming (focused on trust between agents)?

For the brainstorming session, the problem statement above was decomposed into three parts for the three brainstorming rounds. “Formulate criteria and metrics to:”

- [round 1] capture how humans perceive and trust autonomous systems,
- [round 2] characterize methods to measure and modify human trust, and
- [round 3] identify salient features of human decision making that will be relevant to design and operation of trusted autonomous systems.

Theme 2 – Machine-Centric

For this theme, the problem statement used to drive the session was “Formulate criteria and metrics that capture trustworthiness of autonomous systems, characterize methods to measure trustworthiness of autonomous systems, and identify techniques for autonomous systems to dynamically improve trustworthiness.”

Mr. Paul Decker from the U.S. Army Combat Capabilities Development Command, Ground Vehicle Systems Center provided the vision and motivation for the machine-centric aspect of trusted autonomy through a roughly 10-minute presentation of remarks. At the time of the summit, Mr. Decker served as the Deputy Chief Roboticist, where the vision of the Center is to be “The center of excellence for DoD ground vehicle systems modernization and sustainment solutions”, and a stated mission to “develop, integrate, demonstrate, and sustain ground vehicle systems capabilities to support Army modernization priorities and improve readiness.” Additional information about the Army’s Ground Vehicle Systems Center can be found at: https://www.usarmygvsc.com/

For the machine-centric theme, the panel to drive the education and inspiration on this topic consisted of the following members:

- Moderator: Prof. Shaoshuai Mou, Purdue
- DoD Panelist: Dr. Kristin Schaefer-Lay, ARL
- Academic Panelist: Prof. Shreyas Sundaram, Purdue
- VC Panelist: Mr. Dan Madden, Squadra Ventures

The questions posed to the panelist to answer, and expound upon, are listed below.

- How do you introduce metrics to measure the trustworthiness of autonomous systems? Or policies to guide trustworthy autonomy (especially differences from traditional metrics such as robustness against environment uncertainties, resilience against failures/attacks...)
- What do you think are the state-of-art and main research challenges to improve trustworthiness of autonomous systems?
- Any vision on future techniques for autonomous systems to improve trustworthiness? Or new approaches for testing/simulations to verify autonomy with trust.(Especially how classical techniques in control/optimization and recent advance in machine

learning/AI/Data science , or their integration could potentially contribute to solving the challenges.)
- What is your advice for how academia, DoD and VC to work together to address research challenges and also technology commercialization in achieving trustworthiness of autonomous systems?

For the brainstorming session, the problem statement above was decomposed into three parts for the three brainstorming rounds. “Formulate criteria and metrics to:”
- [round 1] capture trustworthiness of autonomous systems,
- [round 2] characterize methods to measure trustworthiness of autonomous systems, and
- [round 3] identify techniques for autonomous systems to dynamically improve trustworthiness.

Theme 3 – Integration-Centric
For this theme, the problem statement used to drive the session was “Formulate criteria and metrics to evaluate the effectiveness of human-machine teams, characterize methods to measure bidirectional trust, and identify closed-loop techniques to dynamically adapt bidirectional trust for improving mission effectiveness.”

Dr. Laura Steckman from Air Force Research Laboratory provided the vision and motivation for the integration-centric aspect of trusted autonomy through a roughly 10-minute presentation of remarks. At the time of the summit, Dr. Steckman was the Program Officer for the Air Force Office of Scientific Research Trust and Influence program. This program does basic research in trust in autonomous systems, socio-digital influence, and computational methods in social science. The program has a goal of advancing the basic understanding of human reliance and teaming, to elucidate how people establish, maintain, and repair trust in agents, both human and machine. In particular, it supports research to build the scientific foundations for designing high-performing, mixed humans-machine teams, through properly calibrated trust. Additional information about Dr. Steckman’s work can be found at: https://www.afrl.af.mil/About-Us/Fact-Sheets/Fact-Sheet-Display/Article/2282109/afosr-information-and-networks/

For the integration-centric theme, the panel to drive the education and inspiration on this topic consisted of the following members:
- Moderator: Prof. Neera Jain, Purdue
- DoD Speaker: Dr. Sandeep Neema, DARPA
- Academic Speaker: Prof. Phil Smith, OSU
- VC Speaker: Ms. Lauren Bedula, Beacon Global Strategies

The questions posed to the panelist to answer, and expound upon, are listed below.
- How do performance metrics defined for humans or machines individually change (or not) when defining metrics for the human-machine team? (or more generally, what makes it difficult to define performance metrics for human-machine teams)?

- The broader research community has long considered human trust in machines. But what does it mean for a machine to trust a human? To what extent is this coupled to a machine's mental models of humans?
- Is it reasonable to expect human-machine teams to emulate human-human ones? If not, how can/should performance metrics be defined to align with reasonable expectations?
- To what extent is human-machine effectiveness context specific and how does that affect the design process we need to mitigate this issue?
- During the design process, how can we evaluate the likely effectiveness of a system design from a human-machine teaming perspective? More narrowly, how can we assess the likely influences of a design on the development of trust and its impacts on performance?
- What is your advice for how academia, DoD and VC can work together to address research challenges and workforce development in achieving effective human-machine teams?

For the brainstorming session, the problem statement was decomposed into three parts for the three brainstorming rounds. "Formulate criteria and metrics to:"

- [round 1] evaluate the effectiveness of human-machine teams,
- [round 2] characterize methods to measure bidirectional trust, and
- [round 3] identify closed-loop techniques to dynamically adapt bidirectional trust for improving mission effectiveness.

D. Design and Implementation Guidelines

Safety and Testing

Category	Need
Human	Human should be able to intervene and override
Technical System	Technical system should have identified safe-to-fail areas
	Technical system should have resilience mechanisms
	The technical system should have a defined state space (must have an Operational Design Domain)
	The technical system should be able to measure and verify that it is operating in proper environmental conditions
Human-Machine Teaming	Humans and machines should be able to detect if/when something goes wrong
Ecosystem	The ecosystem should support system testing in both deployment and simulated environments
	The ecosystem should provide a varying simulated and actual conditions for testing and training that the technical system must operate successfully in
	The ecosystem should provide operational testbeds for system experimentation
	The ecosystem should support real-world/operational experimentation of the system
	The ecosystem should have a simulated environment for experimentation
	The ecosystem should have a testbed for testing and training

Data Collection and Analysis

Category	Need
Technical System	The technical system should be able to do live/automated data collection
	The technical system should be able to assess the current state of the human operator
	The technical system should be able to measure and collect data on human user's state
	The technical system should be able to measure/collect human biometrics
	The technical system should be able to measure usage of features by human
	The technical system should be able to verify data (data verification)
Ecosystem	Ecosystem should collect and have access to metadata of technical system functions and human machine interactions
	The ecosystem should support live/automated data **collection** on human-machine data
	The ecosystem should support live/automated data **analysis** on collected human-machine data

Performance Measurements and Comparisons

Category	Need
Human	Humans should have a clear distinction of mission tasks/conditions/environments
Technical System	Technical systems should be able to measure if tasks are completed 'successfully'
Ecosystem	Ecosystem should enable the measurement / collection the quantity of times it successfully completes the task
	The ecosystem should support technical system performance assessments
	The ecosystem should support measuring and providing comparisons of performance to a predefined standard of performance quality
	The ecosystem should provide pre-established baselines of system performance
	The ecosystem should identify appropriate performance metrics, tailored for the task at hand

Human-Machine Interactions

Category	Need
Human	Humans should observe others using the machine
Human-Machine Teaming	Human and the machine should be able to explain its rationale or actions
	Human and the machine should be able to model each other's behaviors
	Human and machine should be able to understand each other (mutual understanding)
	Human and machine should be able to communicate
	Human and machine should be able to provide feedback in a language the other understands
	Human and machine should be able to know each other’s needs/preferences

	Human and machine should be able to understand level of human usage/reliance
	Human and machine should have adequate feedback between each other
Technical System	The technical system could have “human-like” features
	The technical system should be able to measure/analyze eye/gaze of human
	The technical system should be able to identify and adapt to the user
	The technical system should have adaptive tools for humans to interface with
	The technical system should be able to communicate information over multiple modalities to ensure human and machine have the same understanding of a situation
	The technical system should be able to provide information to human in efficient way
	The technical system should be able to measure decision load and/or cognitive stress of human
Ecosystem	Ecosystem should enable feedback mechanisms between the technical system and human
	Ecosystem should enable technical system co-design and development with the human end user
	Organizing frameworks (e.g., levels of autonomy, clearly defined roles/responsibilities for the human vs. system) should exist within the ecosystem

Transparency and Explainability

Category	Need
Human	Human should be able to identify intentional (e.g., spoofing) tampering of the machine
	Humans should have transparency of the inner workings of machine
Human-Machine Teaming	Human and machine should be able to build mental models of each other
Technical System	The technical system should have the ability to explain decision-making rationale
Ecosystem	Ecosystem should provide appropriate documentation and training about technical system to human users
	Ecosystem should have the ability to identify if technical system explanations were correct

System Confidence and Reliability

Category	Need
Ecosystem	The ecosystem should enable measurement of system error over time
	The ecosystem should support measurement and analysis of safety requirement factors
	The ecosystem should have a taxonomy/criterion for speed-accuracy trade-off
	The ecosystem should support measurement of the number of tasks allocated to a technical system over time
	Ecosystems should be able to provide measurements of reliability and robustness
	The ecosystem should support the identification of model drift
	The ecosystem should support the measurement and calculation of how often the human and machine disagree
	The ecosystem should support the measurement of reliability and robustness of a technical system
	The ecosystem should be able to measure confidence across a human-machine team

Human Trust and Reliability

Category	Need
Technical System	The technical system should be able to keep human in the loop by allocating control to them when necessary
	The technical system should be able to positively identify the specific end user and prove “correctness” of actions and “explain” rationale
Human-Machine Interaction	Human and machine should have a degree of shared control
Ecosystem	The ecosystem should support the measurement and calculation of level of human trust in the system
	The ecosystem should support measurement of varying trust
	The ecosystem should support measurement of trust indications over time

User Training and Adoption

Category	Need
Human	Humans should use the technical system regularly
	Human should have familiarity with the technology and its use
	Human should have and understanding of the task
	Humans should have a willingness to adopt new technology
	Humans should be trained dependent on their experience/role with technology
	Human should be able to adopt technology at rapid rate
Technical System	The technical system should be able to identify level of experience of human
	The technical system should be able to adapt to the experience level of the human
Human-Machine Teaming	Human and machine should be knowledgeable of their limitations

Bias Detection/Mitigation

Category	Need
Technical System	The technical system should be able to identify/detect human biases and mental models
Ecosystem	The ecosystem should be able to aid in identifying bias in both the humans and machine

Adaptability and Flexibility

Category	Need
Human	Human should be able to modify machine features
Technical System	The technical system should be able to modify performance according to human preference
	The technical system should change available tasks as a function of human trust/experience
	The technical system should be able to learn from human interaction
	The technical system should be able to change as a function of complexity
	The technical system should be able to adapt to the trust level of the human
	The technical system should able to analyze how hard a given task must be for a human
Ecosystem	The ecosystem should support standards for the system which are both rigid, yet flexible when appropriate

E. OPTIMA Playbook Meeting Details and Initial Findings

Introduction and Structure

Operational Trust in Mission Autonomy (OPTIMA) is an initiative to deliver trusted Autonomy for Robotic and Intelligent Autonomous Systems in complex, contested missions on the multi-domain battlefield, and enable deployment of effective human-machine teams.

In support of this initiative, a one day working meeting was held at the Johns Hopkins Applied Physics Laboratory, hosted by the Johns Hopkins Institute for Assured Autonomy.

The overarching goal of the gathering was to develop connections between and assemble the objectives, progress, activities, and path forward for each of the key players/efforts across the OPTIMA initiative. An additional goad was to map efforts to key objectives and identify current and future potential gaps.

Structure of the meeting consisted of three unique sections, which are detailed below.

- Setting the Scene:
    - What is the ultimate goal for a final end state or product for each of the key objectives of OPTIMA? [OPTIMA Director shares his vision]
    - Objectives for the working session
- Working session on Key Objective: “Discover and validate quantifiable metrics for trusted autonomy”
    - What is already happening, being done by OPTIMA performers and by others in the general community? [Performers discuss their current efforts and other efforts of which they are aware]
    - Where are the gaps? [Whole group brainstorm session]
    - How do the OPTIMA performers collaborate to help fill those gaps and get to the desired end state/products for OPTIMA? [Collective white boarding session]
- Working session on Key Objective: “Transition frameworks to acquisition life cycle”
    - What is already happening, being done by OPTIMA performers and by others in the general community? [Performers discuss their current efforts and other efforts of which they are aware]
    - Where are the gaps? [Whole group brainstorm session]
    - How do the OPTIMA performers collaborate to help fill those gaps and get to the desired end state/products for OPTIMA? [Collective white boarding session]

Attendees and Stakeholders

Attendees included representatives from ten unique organizations, included all eight organizations who are supporting key initiatives. A draft stakeholder map diagram was generated as a result of this meeting. This diagram consists of each of the ten unique organizations present, a few words description, and how they collaborate with each other, both currently and planned for future collaboration. This diagram can be seen following this section.

A more detailed description can be found below:

OPTIMA Key Activities

Below are all of the activities under the OPTIMA umbrella:

Johns-Hopkins Applied Physics Lab Institute for Assured Autonomy (JHU-APL IAA) to help build out a strategic view and direction of OPTIMA

- Three core questions to investigate for this task:
  - How do you build a body of evidence that a lack of assurance is the limiter from going from proof of concept to scale?
  - How do you generate trusted and trustworthy autonomy?
  - How do you operationalize within the DoD framework?
- What IAA team will do
  - Lead human centered design engagements to gather diverse perspectives, to include moderating discussions of various stakeholders
  - Analyze the data and build into products to support the strategic framework
  - Frame out how to do pilots and proof of concepts engagements/events/activities
- Deliverables
  - Monthly reports, to include specific summaries/reports from each engagement
  - Overall technical report at end of task

Center for Naval Analysis developed the Joint Autonomy Risk Element List in a study entitled, "Dimensions of Autonomous Decision-making", funded by ONR. Operator trust in autonomous systems remains difficult to measure. Can the JAREL be used to develop quantifiable operator trust metrics for autonomous systems?

Approach: CNA will answer this question in a realistic autonomy environment by working with the ONR Expeditionary Robotics program. We will embed a CNA analyst with knowledge of the JAREL into an ONR Expeditionary Robotics program project team. In coordination with the project team the CNA analyst will use the JAREL to define and measure the degree of ethical conformity of an expeditionary robotic system under consideration. This measurement will identify autonomy-related risks that have been mitigated, those that remain present and unmitigated, and risks that can be either introduced or mitigated with minimal effort. This analysis will result in two configurations of the same system, each with a different—and measured—degree of ethical conformity. Each system version will then be provided to operators in the Expeditionary Robotics experimental venue. The operators will be made aware of the ethical conformity risk mitigation steps that have been taken, and which autonomy-related risks remain unmitigated. The CNA analyst will then observe the degree of use of each system version to identify any correlations between the measured degree of a system's ethical conformity, and the degree to which operators trust it enough to use it. This will occur during the experimentation event, and afterward during a hotwash.

Edge Case Research, Inc leads the Autonomy industry in developing the use of assurance cases for safety certification of Autonomy software for self-driving cars. This project will involve drafting, iteratively maintaining, and socializing the results of a study that surveys technologies relevant to the goals of the Operational Trust in Mission Autonomy (OPTIMA) initiative led by the USD(R&E) Director for Autonomy. The study will support OPTIMA goals by engaging across the DOD Autonomy enterprise to outline a strategy and implementation plan for delivering Trusted Autonomy as an operational capability. Tasks performed on the study include, but are not limited to:

- Industry outreach, to include optimizing an approach to most effectively leverage the strengths and constraints of industrial partners, structuring periodic roundtables, technical exchange meetings, and Workshops to foster two-way ideation
- Acquisition Lifecycle Optimization to build innovative pathways to transition from research baselines to frameworks for acquiring Trusted Autonomy, where potential approaches include developing DoD Issuances (DoD Directives, Instructions, etc.); Joint Capabilities Integration and Development System (JCIDS) process adaptations; Test Evaluation Validation & Verification (TEV&V) methods; ongoing Sustainment processes for continuous improvement of assurance of Autonomous Systems

Working with USDR&E and Aerospace Corporation, **Applied Intuition** defines test metrics using the Aerospace Trusted AI (TAI) Framework, develops an appropriate plan to test across the TAI pillars, and builds and runs discrete simulated scenarios that provide pass/fail outcomes that map to the TAI test plan. At a more detailed level, the project roadmap consists of the following steps:

- Define a test plan consisting of specific behavioral competencies that map to the TAI framework. The test plan will include:
  - Define the assurance case
  - Define the behavioral competencies the system must demonstrate
- In the Applied Intuition simulation environment, define the abstract scenario environment in which to execute the test plan, consisting of:
  - **Operational design domain (ODD):** This consists of the vehicle type under test (aerial, ground etc), the geographic location, the environmental conditions, and the sensor models.
  - **Scenario types:** Define test scenarios that will assess components of the TAI. This will involve defining actors, pass/fail criteria, and the acceptable range of test conditions (e.g. weather effects, vehicle speeds).
- Integrate with the Aerospace perception and/or motion planning stack
- Begin scenario testing, iterating across scenario parameters and adjusting models
- Track stack performance over time against the test plan
  - Metrics will track stack performance on a per-scenario basis for model developers, as well as at the Key Performance Indicator level for senior personnel
  - Detailed understanding of where the system is performant and not performant
- Report analyzing the experiment and providing recommendations on further experimentation

Purdue ICON Center - The Summit on Trusted Autonomy Research and Technology (START), held at Purdue University on June 28-29 2022, brought together leading experts from academia, the Department of Defense, and commercialization sectors to brainstorm a research and development roadmap for trusted autonomy and human-autonomy teaming. This summit identified a wide variety of perspectives on the metrics, methods, challenges, and state-of-the art for trusted autonomous systems.

Motivated by the outcomes of START, this project has two deliverables.

1. A survey paper, led by Purdue ICON researchers and their graduate students, with suitable contributions from academic researchers outside ICON, summarizing the current state of the art on trusted autonomous systems. This survey paper will contain a thorough literature review and will be submitted to an appropriate journal for review and publication.
2. A proposal for a special issue in an appropriate journal, to be guest-edited by individuals from Purdue ICON, Johns Hopkins Institute for Assured Autonomy (IAA), and OUSD(R&E) (including Dr. Jaret Riddick), that solicits and collects perspectives or research papers from the community of experts on trusted autonomy and human-autonomy interaction (including the attendees of the START summit).

Aerospace Corporation – approach relies on three concepts

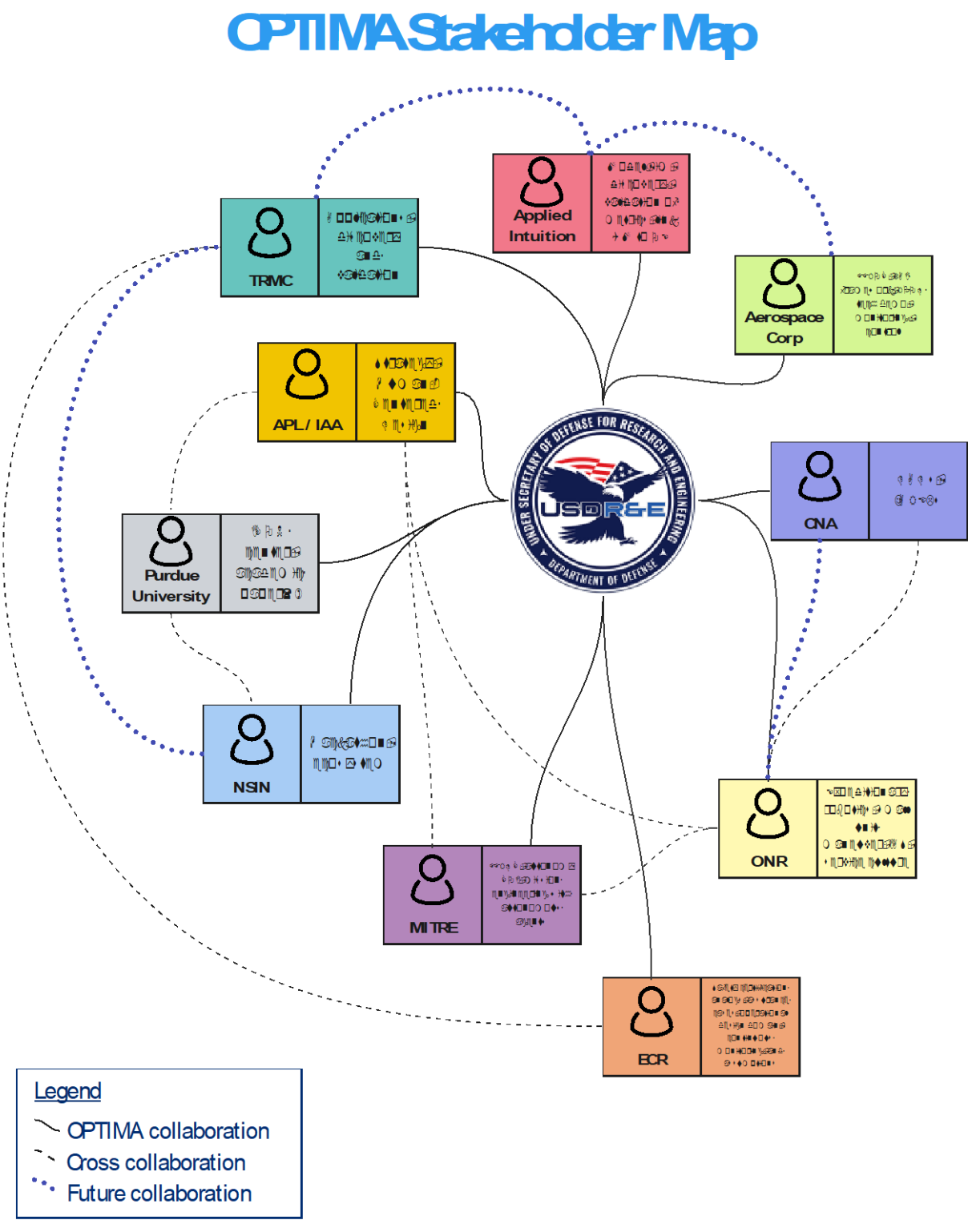
- The Aerospace AI/autonomy Solution Architecting (AASA) process
  - Integrating AI/autonomy from inception should help create faster, cheaper, better solutions
  - Standards don't yet exist for developing AI/autonomy for lifetime trust
  - AASA is a systematic approach to hardware-software co-design across all specialties required in a system, including AI/autonomy
- The Aerospace Trusted AI (TAI) Framework
  - Defines topics required for trusting AI/autonomy in addition to standard performance metrics
  - Organized by topic flow into 4 Threads that define needs, data, trust metrics, & monitoring/control
  - Explicitly included in AASA for development & can apply to already-developed systems
- HMT to accelerate trust
  - Humans mentoring machines to perform technical tasks should be faster & easier than relying solely on humans repeatedly mentoring replacement humans
  - A trained machine teammate can supervise other machine teammates & involve humans as needed
  - HMT can both speed AI/autonomy trust and reduce human operator workload
- These slides discuss these and our modeling, simulation, & analysis effort to quantify benefits of trusted AI

National Security Innovation Network (NSIN) – Integrate OPTIMA principles and outputs into programming linking academic, venture and DoD Mission Partners.

- **Competitions:** Plan and execute a prize competition focused on integrating trusted autonomy into DoD, and identify academic teams and early-stage ventures that have technologies that could serve as proofs of concept for future development. Incorporate labs, PEOs and other stakeholders as part of the process
- **Accelerator:** Integrate early-stage OPTIMA ideas into PACFLT focused accelerator starting in January 2023. Work with other stakeholders within R&E to support testing and evaluation efforts with technologies afterwards
- **Technology Scouting:** Identify teams from prior programs that could be great candidates to pilot OPTIMA principles in testing environments across the Services

TRMC

Stakeholder Map Diagram



Key Objective: Discover and Validate Quantifiable Metrics for Trusted Autonomy

What is already happening, being done by OPTIMA performers, and by others in the general community?

- Unmanned system safety, RCV, ABV, etc.
- Standardization of data collection, testing, and measuring trust
- T&E strategic plan to go to congress in Q1 23
- Develop edge cases, Tier 1 cases that go to first look at modeling trust from assurance perspective
- Experimentation events as opportunity to gather data
- Investigating how risk mitigation improves public acceptance
- AI and autonomy assurance innovation center
- Developing frameworks that include trust
- JIFEX collaboration
- Tech scouting through various programs
- Finding early startups in different technology verticals
- Developing robust roster of teams and companies in the space
- White paper on early summit on trusted autonomy research in autonomy
- "Space Trusted Autonomy Readiness Levels" (https://arxiv.org/abs/2210.09059)
- START summit - data discovery, sifting and sorting through data
- Internal analysis of AI enabled system, gaps and needs
- Program life cycle assessment tied to trust metrics
- ODD taxonomy and requirements definition
- Links to venture capital
- BetterNet - unique approach to assessing confidence and improving reliability of deep AI models used for autonomy
- Definition of OPTIMA objectives and higher res to info paper
- Defining what artifacts are needed to get OPTIMA down the playing field
- Investigating dimensions of decision making (JAREL)
- CAVE Lab - Collaborative & Autonomous Vehicle Ecosystem
- Building testbeds for integrated test and demonstration of autonomy for space applications.
- UL 4600 - safety standard of autonomous products
- Safety performance indicators
- IAS Execution Plan & COAs
- Unmanned task force
- Serious gaming for trust
- OPTIMA definition of success post phase 2 discussion
- Ethical AI - developing tools
- Bias architecture to understand systemic biases
- Research on better methods to engage communities to understand their priorities and needs for assurance, for the dynamic trust element develop scenarios
- Run competitions and hackathons to think about the concepts of how things are developed how risk mitigation improves public acceptance

- ID risks that erode trust
- Policy recommendations
- Re-sim as a powerful way to test against out of distribution data
- Trusted AI
- Startup accelerators, for early-stage ventures looking to get involved with DoD, understanding the business of it
- Center for policy studies
- Human-autonomy interaction workshop at the 2023 American Control Conference currently being planned
- "Trust Case" - tooling to measure
- ONR Scout - attempt to do rapid T&E, potential opportunity
- How to ID manipulated and "uncorrelated" metrics "gamification" of systems
- Navy TF Turing Test - AI assessments, DEVRON model
- DOD domain, working with safety standard MILSTD 882, system safety update for AI/ML
- Links to academia
- Anomaly and detection reporting group human-machine training simulation
- Touring test of human machine teaming
- Ethics policy into engineering requirements
- Modeling and sim group working with programs,
- Trusted AI framework & reference architecture
- Conducting unscripted force on force experiments
- General purpose HMT application, view toward accelerating trust
- Exposing developers to warfighting in 1/4 conditions

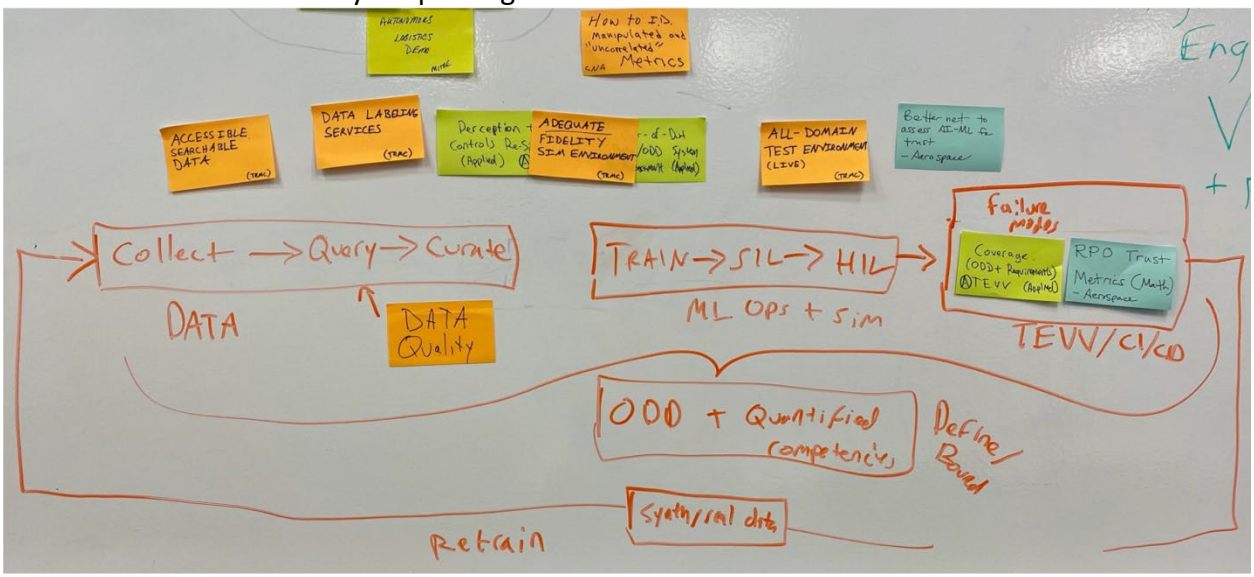
What/Where are the gaps?

- Cyber
- High fidelity simulators
- Research on human-machine teaming and training
- How do we decompose trust into different components?
- Red Teaming
- System AARs

How do the OPTIMA performers collaborate to help fill those gaps and get to the desired end state/products for OPTIMA? (These are brainstormed collaboration ideas or efforts.)

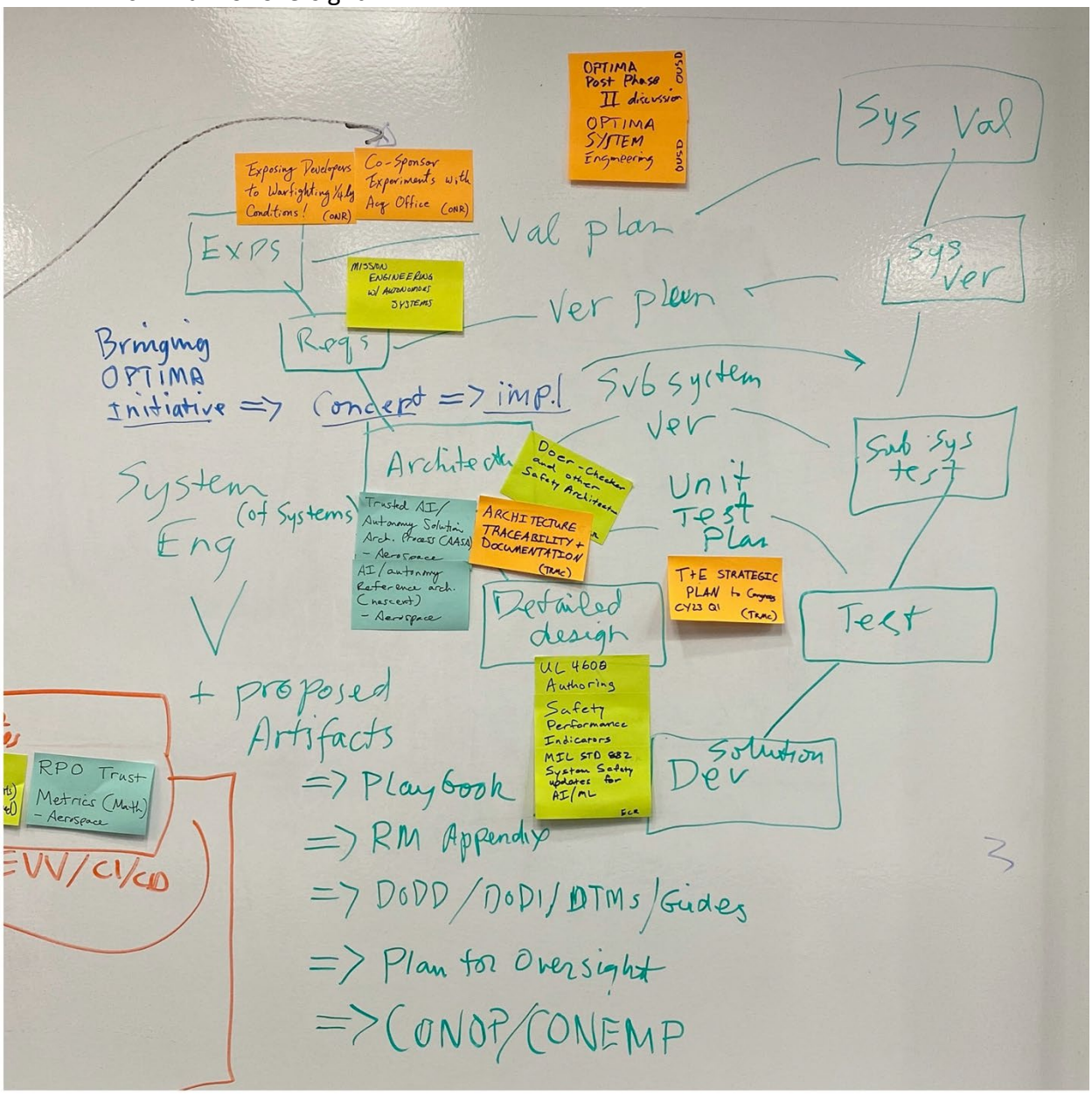
- Technical Solutions into Operational Use
- Leveraging Academic Scholarship
- Assurance use cases
- Post-mission analysis tool
    - Real time display of any sensor
    - Mission replay
    - Data management strategy for post-hoc analysis, applications, etc.
    - Building on an enduring testing ecosystem
- Support of early TRL in autonomy

- Pilot program opportunities
    - ONR experimentation
    - What are we doing today to find the gaps, what it means to them to try and execute it
    - What's needed for it to be a fully autonomous system
- Identifying manipulated Metrics
    - Looking at the work TRMC is doing, scaling what we are doing now, how do we turn it into a production level pipeline
    - Looking at the data, pull in unstructured data that has been collected by autonomous system, curate data sets
    - Generate large unlabeled data set and send off for labeling, build models
    - Sim of adequate fidelity (accurate lighting, environment, etc.)
    - Model development step is very critical. We have decent naval, aerial, ground sim but none of them tied together
    - What happens when you have artificially corrupted data?
        - Bias introduced from collecting different tanks (soviet vs US)
        - Perception riot testing
    - Real world testing, where is the model failing.
        - How do we build synthetic data that starts the process over, and we are always improving



- Modeling and Simulation
- HMT application
    - Envision HMT application that is like the HMT app of the reference architecture.
    - Need IAA's help to do the human machine interface of it
- Testbeds
    - Digital engineering for test best
    - Distributed simulation environments
    - In space testbeds

    - Some of the ONR pieces, the force on force to do the red teaming against what you are doing on the test bed for that
- Systems engineering diagram
    - Phase 1 to 2 (left hand side), 2 to 3 (right hand side)
    - Artifacts like the playbook
    - OPTIMA appendix to roadmaps
    - DODI's
    - Plan for oversight



Key Objective: Transition Frameworks to Acquisition Life Cycle

What is already happening, being done by OPTIMA performers, and by others in the general community?

- QUOKKA (MITRE internal project)
- S&T with an adversary up-front
- R&D with an adversary up-front
- T&E at system level in ecosystem with adversary
- Pull acquisition into military training exercise
- System requirements guidance based on TAI or other frameworks
- System requirements guidance based on AASA process
- De-risking early RL
- Testbed availability
- OPTIMA Playbook
- OPTIMA mission engineering
- OPTIMA in JRASE
- Assurance planning
- Hazard/risk analysis
- Critical function identification
- Level of rigor
- System and human critical function V&V
- Software test completeness

What/Where are the gaps?

- TEV&V requirements processes
- Demos that the processes "work" (faster, cheaper, better systems)
- Human machine team general purpose application that vendors can use
- Knowledge capture and sharing
- Proprietary autonomy
- Post-experiment documentation of resultant hypotheses
- Indicator/metric/proxy library cross-referenced to system, mission, and scenario
- Absent stakeholders
- System definition consensus
- ROM cost estimate for AAI in DoD
- Linking 6.1, 6.2, and 6.3
- Comprehensive SA across programs, activities, and experiments
- Simulation reality gap
- Enduring simulation data environment to explore big ideas in leveraging autonomy
- Coherent enduring tools that build upon previous efforts
- Reference architecture for RFIs, RFPs, etc.
- Open-source software components
- Experimental strategy
- Dedicated AAI penetration testing team
- War simulators do not simulate psychology (e.g., fear)

- Identify the available PPBE/JCIBS/DAS to inject opportunities
- Choosing between deliberate and rapid acquisition authorities
- Disposition analyses to go from rapid to deliberate
- Level of rigor for AI/ML
- Accurate autonomy wargaming
- Continuous feedback loop and capture between operators and acquisition process
- Create a simulation construct to evaluate "big ideas" in trusted autonomy
- Adaption of system engineering "V" for AI-enabled systems to build trust
- UL 4600 - military needs to use it!
- Better communication of needs to industry

How do the OPTIMA performers collaborate to help fill those gaps and get to the desired end state/products for OPTIMA? (This pivoted into asking each attendee stakeholder if they could do one thing for OPTIMA in the next 12 months, what would they do?)

- Think about the human, quantitative metric piece and how to apply it throughout OPTIMA
- Leverage previous Human-Centered Design projects that are analogous, like Toughness, to think through how can we start to share the different sub-components that go into the holistic metric of trust
- Bring everyone together to do a human centered design workshop with an actual assurance case, with actual operators there, warfighters with autonomous systems experience. Really understand what we need to do in learning about trust measurement.
- Work more with early-stage venture and academia
- Would like to see a study on all the existing indicators of trust, and obstacles of getting things measured across the test ranges across the US. What are the possible test metric options?
- Focus on trust-assurance case to define what we mean about trust and to have claims and arguments on trust
- Invite everyone at this meeting to come to an event to see how the current efforts are and how they can be used.
- As an analyst, will make sure whatever measures you choose actually matter and will get you to trusted autonomy (Measurements that matter to whom and to what audience?)
- Given what's been said here, ideally what I would love to happen is to look at the program offices at ONR and get tangible measures of their systems. Everyone does something different, have to figure out for human performance, swarming, etc. What that means for each individual program.
- One quick item to work with everyone to get all these great ideas into an OPTIMA playbook that will take OPTIMA from a concept to a playbook
- Work with aerospace and edgecase to take these metrics into a framework and turn that set of arguments into discrete scenarios; use whatever domain we start with and figure out how to scale across domains
- Work with HMT general application, at least suggest what it should look like, a bunch of open-source pieces to work on, captured things that OPTIMA needs, and writing down

some of the requirements of the trust metric use. Walk through framework and write shall statements.

- I don't think we have an adequate sim engine for the DoD. Do we need to invest in one, I think so. We need a reputable and restricted publication for AI to do the knowledge capture sharing piece.
- Leverage MITRE as a main hub of places to go for understanding who may be working on or already have tools.
- Look at what existing documentation can be rewritten, find a way to rethink how we did TE V&V.