2021 Trust & Influence Program Review – Book of Abstracts

August 16-20, 2021

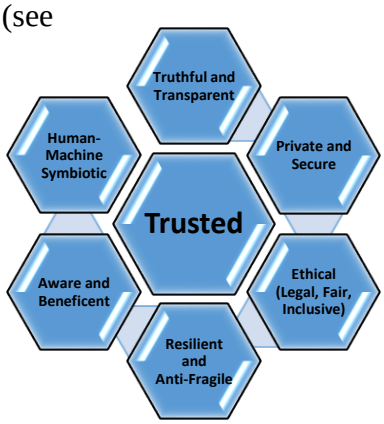
Keynote Speaker

Monday, August 16, 10:00 US Eastern

Trusted Artificial Intelligence

Mark Maybury
Chief Technology Officer
Stanley Black & Decker
1 Constitution Plaza
Hartford, CT
mark.maybury@sbdinc.com
10 June 2021

The promise of artificial intelligence is to make our lives more productive, safe, and joyful. Autonomous systems that can be delegated tasks to perform independently have potential benefits that encompass a broad set of areas including improved governance, defense, transportation, energy, education, and healthcare. Realizing progress across these domains, however, requires overcoming a number of fundamental challenges (see Figure) including:



Trusted Autonomy – Ability of humans and institutions to govern, certify, control, and evolve autonomous systems to assure their appropriate operation in society. There is need for "certifiable trust" via verification and validation (DSB 2012) of near infinite state, evolving learned machines. Trust occurs when trusted (secure and safe) agents become trustworthy (authorized with delegable tasks with degrees of independence) because of performance and anti-fragility in a range of situations and environments.

Truthful and Transparent – A systems perceptions, decision and actions are (as much as possible) accurate, observable, predictable, directable (Johnson et al. 2014) and auditable. This includes mutual understanding of shared goals, self-evident or perspicuously explained rational for decisions that ensure harmony between machine architectures and human mental models of system performance and competence, overcoming any fundamentally different methods and bases of reasoning between human-machines.

Human-Machine Symbiosis – Improvements in the degree of autonomy (in, on, and out of the loop) in a range of increasingly challenging tasks and contexts as well as the appropriate and effective distribution of work across humans and machine, fostering human reskilling and upskilling. Intuitive engagement requires advancement of intelligent multimodal and possibly multiparty conversations to support effective cognitive assistance.

Self and Environmentally Aware and Beneficent – Improved methods for intelligent sensing, situational and self-awareness, and agile response in a range of physical and cognitive assistance situations. They aim to do no harm and be as cooperative and supportive as possible.

Private and Secure – Improved methods for intelligent authentication, situational awareness of threats, and autonomous counter in contested environments.

Resilient and Anti-Fragile – Whereas resilience is the ability to operate through a variety of contested environments by deflecting, absorbing or recovering from insults, anti-fragility enables systems to evolve (physically and cognitively) to thrive in extreme environments through agility, diversity, and evolution. For example, machine learning data diversity, V&V, and stress testing are essential (DSB 2020), challenging because of non-deterministic inputs, lack of repeatability and emergent behavior given machine learning.

Ethical and Inclusive – the evolution of regulation (laws and judicial systems) to address the increasing roles and responsibilities of increasingly prevalent and sophisticated AI. Ensuring access (e.g., to knowledge and services), fairness (of algorithms and services), and equity fostering social inclusivity across diverse cognitive, ethnic and identity boundaries.

While seemingly dauting, public private partnering across industry, academia, government, and non-profits promises rapid community progress for what is an exciting future which promises to advance both human and machine intelligence.

References

Dahm, W. (ed). 15 May 2010. *Technology Horizons*, Volume 1; AF/ST-TR-10-01;

David, R. and Nielsen, P. (co-chairs). 2015. Report of the Defense Science Board Task Force on Autonomy

Johnson, M., Bradshaw, J.M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B. and Sierhuis, M. 2014. "Coactive Design: Designing Support for Interdependence in Joint Activity", *Journal of Human-Robot Interaction*, 3 (1): 43-69.

Klein, Gary, et al. 2004. Ten Challenges for Making Automation a 'Team Player' in Joint Human Agent Activity. *IEEE Intelligent Systems 19* (6): 91–95.

Koch, M., Manuylov, I., and Smolka, M. 2019. Robots and firms. *Centre for Economic Policy Research.* voxeu.org/article/robots-and-firms

Maybury, M. T. 2020. AI and Manufacturing Roadmap. AAAI Spring Symposium on AI and Manufacturing. March 23-25, Stanford, CA.

Maybury, M. and Carlini, J. (co-chairs). 2020. Report of the Defense Science Board Task Force on Counter Autonomy.

Murphy, R. and Shields, J. (co-chairs). 2012. Report of the Defense Science Board Task Force on the The Role of Autonomy in DoD Systems.

WEF 2019. Responsible Use of Technology. April 2019. World Economic Forum.

WEF 2021. Artificial Intelligence Ethics Framework. Feb 2021. World Economic Forum.

Project Key:
T – Trust
I – Influence
M – MINERVA
S – Science of Science-MINERVA
D –Defense Education and Civilian University Research (DECUR) Partnership - MINERVA
DR – Defense University Research Instrumentation Program (DURIP)
TAI – Trusted AI Challenge Winner
Y – Young Investigator Program (YIP)
SRC – Space Research Challenge Winner
X – International Program Research

Alphabetical by Principal Investigator

[T] Dr. Gene Alarcon, 711 Human Performance Wing (711th HPW) (AFRL/RH)

Automation Biases in Human-Robot Trust Interactions

The Automation Biases in Human Robot Trust Interactions is to determine the differences in how humans trust people as compared to how humans trust robots. We have utilized experimental manipulations to determine differences in human-human trust and human-robot trust. It is hypothesized that humans will have higher perceptions of other human's trustworthiness and will perform more trusting behaviors with other humans than with robots. It is also hypothesized that performance will have a greater impact on changes in perceptions of trustworthiness and trusting behavior in human-robot teaming. Results of the current research indicates there are differences between human-human and human-robot teaming. In particular, some of these differences were moderated by perfect automation schemas. Lastly, we also discuss the influence of anthropomorphism on the trustworthiness perceptions and trust behaviors of in robots.

[S, M] Dr. YY Ahn, Indiana University

Science Genome: A Scholarly Graph Embedding Framework to Uncover the Fundamental Dynamics of Scientific Enterprise

Science genome project aims to develop, understand, and apply representation learning methods for modeling the dynamics of scientific enterprise, which is captured in complex heterogeneous networks of scientists, papers, journals, ideas, and others. This presentation will provide a brief overview of the project and showcase some of the recent findings about the power of embedding approach.

[M, I] Dr. Scott Atran, Changing Character of War Centre, Pembroke College, University of Oxford

Addressing Resilience in the Western Alliance Against Fragmentation: Willingness to Sacrifice and the Spiritual Dimension of Intergroup Cooperation and Conflict

A main goal of this project is to determine how shifting value priorities and group attachments in the Western alliance can be best configured to bolster resistance to internal fragmentation and key external threats. In this presentation we focus on malign influence operations by Russia, with respect to Covid-19, and China, regarding bilateral and multilateral relations with Europe, to undermine the Western alliance so that we can better identify: 1. which values, issues and groups these adversaries prioritize for targeting and manipulation, 2. how effective is such prioritization and targeting, and 3. what steps might effectively counter such malign information operations. Broadly, whereas Russia focuses on negative messaging on issues that affect core cultural values related to family, community and country, China concentrates on positive messaging that emphasizes cooperation on material security (e.g., economy, environment, territorial sovereignty) in opposition to US-instigated disregard or subversion of such putative cooperative endeavor. Finally, we discusses recent changes in adversaries' approach to malign information operations against the Western alliance and suggest ways of countering.

[I] Dr. Scott Atran, Artis International

Leveraging Wedge Issues Among Populations Susceptible to the Influence or Control of The Islamic State (and other radical jihadist groups)

In this project we explore: 1. Which Psycho-Social Dimensions are most apt to enhance or degrade radical Islamist and anti-Western influence—for example, Material/Military vs. Spiritual/Sacred—and 2. What issues within these dimensions are most apt to enhance or degrade influence. Here we focus on a joint effort by ARTIS and the US Air Force Academy to investigate (1) followed by a brief account of ongoing work concerning (2). Across eight studies in four different countries (n = 2,556), we extend insights from field investigations in conflict zones to offline and online surveys, showing that personal spiritual formidability—an individual's inner strength and character—is positively associated with the will to fight and sacrifice for others. The physical formidability of groups in conflict has long been credited as the primary factor in decisions to fight or flee during conflict. Our studies in Iraq, Palestine, and Morocco reveal that personal spiritual formidability is more strongly associated with willingness to fight and make costly self-sacrifices for a primary reference group than physical formidability. A follow-on study among US Air Force Academy cadets further suggests that this effect is mediated by a stronger loyalty to the group, a finding replicated in a separate study with a different sample. Results indicate that personal spiritual formidability is a primary determinant of the will to fight across cultures; and this individual-level factor, propelled by loyal bonds disposes combatants as well as ordinary citizens to fight at great personal risk. Finally, in ongoing work, we probe issues affecting perceptions of spiritual formidability as a force in anti-Western sentiment in Turkey and the UK: for example, how recent fighting in Gaza has shaped Muslim perceptions that Hamas achieved spiritual victory despite material loss against Israeli forces, and how these perceptions affect allegiances and likely behaviors toward the US, NATO, jihadi groups, democratic values, and more.

[M, I] Dr. Noémie Bouhana, University College London

The Social Ecology of Radicalisation

In recent years, research on radicalization has privileged the investigation of individual-level attributes. Attention to the immediate socio-physical environment in which radicalization takes place and to the processes through which individuals interact with this environment has been less systematic, even though interest in the "where" of extremism is steadily growing. This presentation synthesises the findings of the Minerva SER project, drawing from social ecological surveys conducted in London, analyses of radicalization hotspots in the UK, and interviews with local youths, former extremists and CVE practitioners carried out in Aarhus, Belfast and London. It highlights, notably, how a developing knowledge of contextual mechanisms should lead us to reconsider the *de facto* distinction between 'online' and 'offline' extremist settings, making the case that we should reframe our inquiries within a more integrated, 'onlife' perspective. It concludes with a brief overview of follow-up projects funded by the UK Home Office with a view of operationalizing the environmental assessment of extremism risk through a framework developed by the Principal Investigator known as S5.

[T, X] Angelo Cangelosi, Wenxuan Mou, Samuele Vinanzi, University of Manchester

Artificial Theory of Mind in Human-Robot Interaction

Human social dynamics rely upon the ability to correctly attribute beliefs and goals to other people. This set of abilities have been collectively called Theory of Mind (ToM). This allows us to understand the actions, beliefs and intentions of others within an intentional or goal-directed framework.

Inspired by this, we construct a ToM computational model using deep neural networks, so that the ToM can be instantized on a robot. Extensive experiments were conducted with a humanoid robot iCub in simulation in a game scenario using Pybullet. We found that a robot can be trained to develop ToM ability after observing iCub playing the game. The ToM ability of the robot work both for predicting the actions of iCub and for understanding false-beliefs.

In order to investigate intention reading and ToM in hybrid multi-agent scenarios, we designed a cognitive architecture that allows a robot to observe its environment and extrapolate qualitative spatial relationships (QSRs) between agents and objects. Sequences of QSRs are translated into human motions, then actions and finally probabilistic goals. This architecture, plus a communication and arbitration mechanism, will be used by the robots, in a simulated multi-robot scenario, to infer the team goal of the humans and to collaborate accordingly.

In addition, we also investigated whether the perception of ToM abilities on a robotic agent influences human-robot trust over time in an iterative game scenario. To this end, participants played an Investment Game with a humanoid robot (Pepper) that was presented as having either low-level or high-level ToM. During the game, the participants were asked to pick a sum of money to invest in the robot. The amount invested was used as the main measurement of human-robot trust. Our experimental results show that robots possessing a high-level of ToM abilities were trusted more than the robots presented with low-level ToM skills.

[T] Dr. Meredith Carroll, Dr. Jessica Wildman, Dr. Amanda Thayer, Florida Institute of Technology

Trust Dynamics in Heterogeneous Human-Agent Teams: Applying Multilevel and Unobtrusive Perspectives

This effort is a 3-year, multidisciplinary, collaborative effort being conducted by Florida Tech's Advancing Technology-interaction and Learning in Aviation Systems (ATLAS) Lab and Institute for Culture, Collaboration, and Management (ICCM) to develop and begin to validate a Multilevel, Dynamic Framework of Trust Dynamics in Human-Agent Teams (HATs) and associated unobtrusive measures of HAT trust. The first year of the effort will focus on conceptualizing a theoretical Multilevel, Dynamic Framework of Trust in Human-Agent Teams. Building on the extensive expertise of the current team in both the HAT and Human-Human Team literature, we will review recent progress in order to inform development of a theoretical model that will be validated in later experimentation. Years 2 and 3 will include conducting a series of experiments designed to validate aspects of the model such as the influence of different types of trust violations and repair strategies in teams, and how various compilational patterns of these events across teammates impacts trust. We will utilize one to two experimental testbeds, including the Multi-UAV Simulator for Teaming Research and Evaluation of Autonomous Missions (STREAM), a testbed for multi-UAV operations developed in collaboration with the Air Force Research Lab's Gaming Research Integration for Learning Laboratory that simulates multiple autonomous UAVs with decision making capabilities in ISR missions. We will also potentially utilize ARMA 3, a commercial-off-the-shelf military combat simulator that includes virtual, autonomous agents which act as non-player controlled soldiers. It is anticipated that this effort will result in an empirically validated (a) model of HAT Trust Dynamics and (b) unobtrusive and objective measures of HAT trust. The outcomes will help inform the design, implementation and training of HAT agents and human teammates resulting in improved team dynamics, and increased team performance and mission effectiveness.

[T, I] Dr. Jaime Banks, Texas Tech University (Dr. Duncan Lorimer, West Virginia University)

Moral Agency in Robot-Human Interactions (MARIA): Perceptions, Trust, and Influence

The MARIA Project examines how perceptions of social robots as moral agents contributes to the ways they may be trusted and influential. After Year 1 of the project focused on non/conscious processes associated with mind perception and moral intuition and Year 2 focused on boundary conditions for perceived moral agency's (PMA) promotion of trust, this current third year's work focuses on PMA's influence on implicit and explicit forms of influence through five investigations. (1) PMA's moral-standing counterpart—perceived moral patiency (PMP), that robots are deserving of moral consideration—was inductively mapped, finding 36 key forms of PMP. That mapping served as the foundation for (2) validation of a PMP scale and experimental stimuli for the following projects. (3) A study of PMA/PMP relationship was tested for adherence to Moral Typecasting Theory (predicting an asymmetrical, inverse relation) and for PMP's promotion of trust; the predicted inverse relation was not evident, as we instead observed a positive association between PMP and both PMA and trust. By way of influence, (4) data show that the presence of a social robot observer elicits no task-performance improvements or degradations (i.e., implicit influence) compared to human observer or performing alone, and (5) a final study still in progress

will examine whether PMA may influence the extent to which a person may be persuaded by a robot (i.e., explicit influence).

[TAI, T] Professor Peter Bruza, Queensland University of Technology, Australia

A quantum holistic trust model for calibrating trusted interactions between a human agent and AI system.

The aim of this presentation is to provide an overview of theoretical background, methodology and objectives of our project which is planned to begin Q4, 2021.The broad objective of is to model the gap between a human agent's expectation of an AI system and their perceived trust while interacting with the system. This will be driven by novel research into interactive trust, stemming from theories within the field of quantum cognition, with the aim of producing innovative theory-backed approaches. A simulated AI prototype system will assist a human decision-maker in a decision scenario. The simulation will be programmed to display specific configurations of information based on the scenario, together with trust expectation audits. Data collected from the audits will prime the quantum holistic model of trust expectation, the aim of which is to capture the agent holistically sensing trust across reliability and benevolence. The outcome of this project is an empirically tested theoretical model for the calibration of dynamically changing expectations of human and AI system.

[TAI, T] Dr. Theodora Chaspari, Texas A&M University

Investigating human trust in AI: A case-study of human-AI collaboration on a speech-based data analytics task

Abstract: This talk will discuss trust in artificial intelligence (AI) during a human-in-the-loop collaborative speech-based data analytics task (DAT). The increasing use of AI in the military gives rise to questions about how much AI systems are and can be trusted by their users. There is evidence suggesting that the dimensions of trust in automation and trust in AI are common. However, despite the similarities, it is unclear whether trust within human-AI collaboration actually has the same dynamic as in the human-machine context (e.g., vehicles), therefore this talk will discuss ways for obtaining an improved understanding as to how human trust is established, maintained, and repaired when teaming up with AI systems. Target dimensions of trust include trust calibration, resolution, and specificity. Individual (i.e., annotators' expertise, personality, trust propensity) and system-based (i.e., AI performance and explainability) characteristics will be presented as potential factors that affect human trust in the AI. The discussion will be exemplified via a speech-based DAT, in which human users are called to collaborate with an AI system in detecting deceptive/truthful speech, a challenging DAT of high relevance to the military.

[T, X] Prof Antonio Chella, University of Palermo, Italy
RESPECT: Robot innEr SPEeCH for Trust

The project RESPECT explores the strategic coupling of cognitive robotics modeling and empirical human-robot interaction experiments to analyze the role of robot inner speech in the development of trust interactions between humans and robots.

During the first year, a computational model of inner speech was developed and validated. During this second year, the team project analyzed the relationships between robot inner speech and trustworthy interactions between humans and robots, proposing robot inner speech as a new critic feature to be considered in human-robot interactions.

Then, the team members refined and extended the computational model of inner speech developed during the first year. An implementation of the model was proposed using the ACT-R cognitive architecture. The model was linked to the robot Pepper using ROS and validated using current software standards.

Finally, the project team began the empirical experiments on human-robot interactions employing robot inner speech. Preliminary results confirmed that robot inner speech is a significant feature in trustworthy human-robot interactions.

Notably, the goals and scope of the RESPECT project generated a great deal of media attention.

[T] Dr. Nancy Cooke, Arizona State University

Improving Situation Awareness in Distributed Human-Robot Teams

In order to facilitate team situation awareness, trust, and performance in human-robot teams we ask the following research questions: What are the essential cognitive characteristics of robots required for team situation awareness and effective teaming in distributed human-robot teams? What representations and algorithms are needed to enable robots to realize such behaviors? This research is accomplished by a multidisciplinary team from human systems engineering and computer science. In this talk we will present our empirical and modeling efforts over the last year.

[T, SRC] Dr. Nancy Cooke, Arizona State University

Trusted Distributed Human-Machine Teaming for Safe and Effective Space-based Missions

A challenge of space-based missions is effective teaming in a geographically and temporally (i.e., spatio-temporal) distributed environment. The geographic distribution of teammates coupled with the variable communication latency challenges effective teamwork. This challenge is exacerbated by the team complexity in a heterogeneous multiteam system composed of humans, robots, and Artificial Intelligent (AI) agents. The long-term objective of this research is to develop an AI agent that monitors distributed human machine teams (HMTs) in space-based missions to identify potential team states (e.g., fatigued, conflict, trust) and intervene when needed to improve teamwork and team effectiveness. In the first half of the first year we have identified challenges from experts in space operations, developed a scenario to reflect those challenges, and identified sensor data for HMT monitoring.

[M, T] Dr. Ewart de Visser, Warfighter Effectiveness Research Center, US Air Force Academy, and Dr. Frank Krueger, George Mason University

Examining oxytocin as a causal mechanism for long-term bonding between humans and autonomy

The overall objective of this work is to examine oxytocin (OT) as a causal mechanism for long-term bonding in human-autonomy dyads. We define bonding as the development of a close friendship, strong comradery, or deep affiliation between members of the same group and in this work, we will use the terms bonding and affiliation interchangeably. Our central hypothesis is that autonomous systems (virtual agents, and social robots) equipped with features optimized for social interaction will engage the human social affiliation system and that OT is a causal factor for this affiliative behavior. By adopting a comprehensive, interdisciplinary research program drawing on psychology, ergonomics, and neuroscience, we are evaluating our central hypothesis by testing 1) the feasibility of short-term affiliation in human-autonomy dyads and 2) the feasibility of long-term affiliation in human-autonomy dyads. We have developed protocols for both short and long-term affiliation and experimental evaluation is currently in progress. In particular, at Auburn University, we have developed a long-term protocol to compare human fMRI and OT responses in human-dog affiliation and human-robot affiliation comparing interactions with real dogs to Sony's Aibo robotic dog. We also present preliminary data collected at Drexel University examining short-term affiliation with a humanoid robot using a rapport-building task. Functional NIRS data revealed that the hemodynamic brain activity over the right medial prefrontal cortex region was significantly higher during error-free attitude interaction with the robot indicating that participants were more socially engaged in this condition. Overall, this work demonstrates that the assessment of social cognition via neurocognitive measures is a promising approach to advance the human-robot interaction field by providing insights from human cognitive processes during human-robot interactions.

[M, I] Dr. Brian Ekdale, University of Iowa

Algorithmic Personalization and Online Radicalization: A Mixed Methods Approach

This project aims, first, to identify the technological, psychological, and cultural factors that contribute to radicalization of vulnerable populations on social media and, second, to develop tools that can identify and predict radicalization on social media. Our mixed methods approach includes a longitudinal survey, behavioral data tracking, qualitative interviews, and the design of a predictive algorithm to identify communities vulnerable to future exposure to extremist content online. Each of these research methods will produce valuable findings that will help explain the relationship between algorithmic personalization and online radicalization.

[M, T] Dr. Nicholas G. Evans, University of Massachusetts – Lowell

Neil D. Shortland, Jonathan D. Moreno, Michael L. Gross, Blake Hereth

The ethics of warfighter participation in the development and testing of AI-driven performance enhancements

The use of artificial intelligence, combined with emerging neuroscience, is purported to be a critical element in national security and warfighter preparedness in future conflicts. On the one hand, next-gen devices will function as therapeutic interventions, e.g. BCIs controlling prosthetics to restore capacity, including restoring nervous system feedback through artificial limbs. However, these devices can also maintain, and **enhance** human performance during training and deployment. What is not determined, however, are the conditions under which these performance enhancements ought to be tested on, or used by, warfighters.

In this panel we report on our current project that poses the following questions:

What ethical considerations inform the role of warfighters as research subjects, and as users of experimental neurotechnologies?

The proposed project will answer these questions by completing two specific aims:

1) Identify special ethical features of warfighters as research subjects as they pertain to the testing and use of these technologies for restoring, maintaining and/or improving human performance and capacity;
2) Investigate how military personnel weigh risks and benefits of deploying experimental technologies.

We report on the status of the project, ongoing elements of the research, new collaborations, and how this project will connect to the wider Minerva Research Initiative ecosystem.

[M, D] Dr. Nina Fefferman, University of Tennessee – Knoxville

A Taxonomy of Communication Functions on Higher-Order Topologies

Abstract: Formalizing the mathematics of multi-domain systems involves understanding the impact of structure in communication among individuals. Truly simple spread processes rely on characterizations of topology alone (e.g. as in the case of information percolation in a random network relying only on a threshold in network edge density), however, as the type of communication increases in complexity, so too does the nature of information needed to characterize the spread of information/understanding. Most often, each particular case (i.e. communication function) is studied as its own object. For example, diffusion and token passing are the subjects of different studies, explored by both simulation and development of analytic theory. In this talk, we instead propose a taxonomy for the classification of communication functions on topological structures (simplicial sets) that are likely to produce similar outcomes in the success of communication based on categorical feature sets. We will walk through some concrete examples of communication functions and explore how and why these feature sets may be meaningful drivers of their behavior across higher-order topologies of interaction.

[T, X] Dr. Luciana Ferrer, Fundación Ciencias Exactas y Naturales, Argentina

Detecting Distrust Towards a Virtual Assistant from the User's Speech

Research has shown that trust is an essential aspect of human-computer interaction directly determining the degree to which the person is willing to use a system. An automatic prediction of the level of trust that a user has on a certain system could be used to attempt to correct potential distrust

by having the system take relevant actions like, for example, apologizing or explaining its decisions. In our work, we explore the feasibility of automatically detecting the level of trust that a user has towards a virtual assistant (VA) based on their speech. Since, to our knowledge, no public databases were available to study this problem, we developed a novel protocol for collecting speech data from subjects induced to have different degrees of trust in the skills of a VA. Using the resulting dataset, we trained an automatic system to detect a proxy for the user's trust toward the VA's abilities. This system resulted in an accuracy of up to 76%, compared to a random baseline of 50%, indicating that the trust level can be estimated from the speech of the user. These results, while encouraging, should be considered preliminary given the restricted nature of the collection protocol. In the next phase of our project we plan to collect a new dataset with a more ambitious scenario which will elicit more variable and rich speech. We will use this new dataset to test our proposed system, compare conclusions with those obtained on the previous dataset, and develop novel approaches for the automatic detection of trust. Further, the new dataset will allow us to study the issue of teaming during human-computer interaction.

[I] Daniel M.T. Fessler, University of California – Los Angeles

Emotions, Attitudes, and the Moral Marketplace: The Psychological, Social, and Situational Determinants of Prosociality and Antisociality

The presence of altruistic individuals enhances the payoffs, and reduces the risks, of prosociality. Third-party witnesses to altruism experience an uplifting emotion, termed *elevation*, which impels them to behave prosocially. However, because others' motives can only be inferred, the extent to which altruistic acts elicit elevation is contingent on prior beliefs; these form a lens through which the witness interprets acts of apparent altruism. *Idealists,* who expect others to behave cooperatively, are susceptible to elevation, whereas *cynics,* who expect the opposite, are not. Within-group prosociality often occurs in the context of inter-group conflict. Leveraging U.S. social movements of the past year, we demonstrate that idealism can be contingent on coalitional affiliation, thus focusing elevation-motivated altruism on the welfare of comrades to the exclusion of opponents.

[T] Dr. Gregory Funke, 711 Human Performance Wing (711th HPW) (AFRL/RH)

Working with or around you? Investigating team growth and adaptation with a cooperative machine agent

Strategic research guidance from the U.S. Department of Defense (DoD) and U.S. Air Force states a need for advanced autonomous/agent systems capable of human-machine teaming (HMT) to meet critical Air Force objectives. The goal is to eventually develop sophisticated machine agents that act as full teammates, able to contribute to team planning and capable of executing complex tasks with minimal human oversight.

Present-day agents are largely unable to meet these Department of Defense (DoD) objectives, in part because of limitations in their ability to interpret others' intentions to and communicate effectively. Until fully autonomous agents with these capabilities are realized, it is likely that the DOD will rely on semi-autonomous agent systems to partially fulfill the need for HMT. These agents are likely to have the capability to handle narrow unknowns and only a limited ability to communicate and aid their human

teammates in critical aspects of teamwork, such as the development and adaptation of team strategies. Regardless of these limitations, agents will need to act interdependently with teammates in pursuit of common goals, which in turn will hinge on the ability of the agent teammate to cooperate by adapting its tactics to support the strategic direction of the team. The inclusion of semi-autonomous teammates with different capacities to cooperate in this manner is likely to have an impact on critical team processes, such as the formation of shared mental models (functional understanding held in common by teammates), team communications, and team trust, all of which will have consequences for team performance. However, the ways these effects will manifest in the context of HMT is largely unknown. Furthermore, the nature of these processes are likely to evolve as human teammates acquire a greater understanding of an agent's capabilities and limitations.

In response to these gaps, our research approach in this project includes a manipulation of a machine agent's capacity to cooperate (CTC) with its human teammates and a manipulation of team experience. We will examine the effects of these factors on team performance and critical team processes, such as communication, shared mental models, and trust, in a computerized version of the complex and strategically oriented board game Pandemic. Our efforts will include a longitudinal design where we will collect behavioral and self-reported data from teams of one human and three machine agents as they play Pandemic over the course of several experimental sessions.

[M, I] Dr. Erik Gartzke, University of California – San Diego

Complex Linkages, Ambivalent Ties: Global Security and Economic Interdependence in the 21st Century

The international community faces a growing set of rapidly evolving, dauntingly complex political challenges related to economic interdependence and security. A revisionist Russia has unilaterally redrawn the borders of Europe, and increasingly seeks to confront the United States and its allies in nontraditional domains. Meanwhile, the economic and military rise of China threatens to reshape the global order for the first time in decades. These dynamics and others suggest a return to traditional Great Power competition. However, they are also occurring in a historically novel context. Unprecedented levels of economic interdependence and the impractical nature of major war among nuclear powers stand as twin pillars discouraging some types of aggression, while enabling others. Economic interdependence makes military conflict between great powers more costly at the same time that it facilitates novel forms of economic coercion, ranging from the raising of tariffs and embargoing of strategic resources, to blacklisting technology firms with close relationships to governments (Huawei) and targeting state and non-state actors through their dependence on financial networks (Al Qaeda, Iran). Interdependence generated by the revolutions in information technology have also simultaneously produced tremendous economic efficiencies and rising vulnerabilities to new forms of aggression (backdoors in telecommunications infrastructure, ransomware attacks on municipal governments, election interference, corporate espionage). Keeping these complex interdependencies and their potential for conflict in mind, we seek to answer the following questions: 1. How do competition and conflict function in the presence of dense economic ties? 2. How can the U.S. best use economic power to achieve its national interests while avoiding armed conflict? 3. Under what conditions is economic coercion most effective? The complexity of economic interdependence ensures contrasting consequences in different contexts. While interdependence reduces the economic incentives of states to conquer their neighbors' territory outright, it may increase incentives for nations

to demand changes to a neighbor's policies or politics, including antitrust and non-competitive behavior. Trade and foreign investment allow states to gain a great deal of wealth from their neighbors without conquering them. Conversely, it is precisely because the United States is so interdependent with other nations that US leaders have such a strong interest in Chinese industrial policy, monetary policy, antitrust regulations, intellectual property rights, and Chinese relationships with other nations. Interdependence thus both increases the economic costs of large-scale interstate war (e.g. Gartzke 2007) and also increases the range of tools available to nations to coercive one another short of military violence. At the same time, growing reliance on a global technical infrastructure facilitates new forms of conflict such as cyber espionage, influence, and subversion (Lindsay forthcoming). These complexities and offsetting effects yield strategic realities that appear at first to be counterintuitive. For example, states may utilize low intensity military violence as a substitute for economic confrontation where commercial relations are profitable and fragile (Zhang 2018). Gartzke, UC San Diego 4 We propose to create new knowledge in three related areas: 1. Identify asymmetric economic ties that can most readily be exploited in coercive diplomacy and the conditions under which economic coercion is most effective; 2. Assess the ways in which economic solvency, ally cooperation, firm compliance and technological context work to shape states' power and exposure to economic coercion; 3. Detail the conditions under which interdependence can change the nature of military conflict in terms of entry, frequency, escalation, intensity, and duration. In all three of these areas, making informed policy recommendations based on existing research is hampered by both theoretical disarray and a lack of appropriate measures of interdependence. The proposed project will begin by disaggregating economic interdependence in order to identify those types of relationships that are most useful as tools of economic statecraft. We will combine these new measures with data on the other variables that do the most to shape the effectiveness of economic tools of influence, including ally cooperation, firm compliance, economic solvency, and the vulnerability of internet infrastructure. These new data will enable the precise analysis needed to identify key economic sources and processes of power and vulnerability to untangle the relationship between economic interdependence and military conflict. In addition to conducting our own analysis, we will make these data immediately available to researchers both inside and outside of government, in effect facilitating the rapid production of actionable knowledge.

[M, D] Dr. Erik Gartzke, University of California – San Diego and Dr. David Sacko, US Air Force Academy

Economic Interdependence and National Security in the 21st Century

An unprecedented level of economic interdependence complicates development of any U.S. strategy for competition with rivals like China or Russia. This heightened economic interdependence between allies and competitors alike both shapes the costs of military conflict and makes available new tools of economic statecraft and coercion. How can the United States and its allies strategically manage their common commercial ties to avoid vulnerabilities and achieve leverage against strategic competitors? How can the United States coordinate use of economic power with its strategic partners to achieve key interests while minimizing the likelihood of armed conflict, or at the least, minimizing the likelihood of American casualties? More broadly, to what extent and under what conditions can tools of economic statecraft supplement or even supplant military power and reduce lethal risk to U.S. military personnel? Our research aims to answer these questions by creating new insights focused on three areas: (1) exploring the ways in which asymmetric interdependence in investments and financial flows shape

states' power and vulnerability with respect to economic coercion; (2) defining the ways interdependence complicates cooperation and coordination among allies; and (3) creating a country-specific decision-guidance framework to help defense policy makers evaluate economic statecraft tools with respect to China and Russia. Additionally, a key component of the proposed DECUR project is engaging civilian and military students through the application of new experiential learning techniques deployed across all five of the academic and Professional Military Education (PME) institutions joining in this proposal. Students and cadets will work and train collaboratively across institutions, both developing cutting-edge data science skills and applying those skills to contribute to the substantive research at the core of this proposal.

This project builds on extensive data science work already being conducted by our team with respect to developing sophisticated new indicators of mutual and asymmetric dependence and leverages these new measures to explore different, more policy relevant questions than has been the case with research in the past. In particular, we seek to elicit effective tools and strategies for economic statecraft. Our proposed research complements both previously funded Minerva work conducted at cPASS (award #'s N00014-14-1-0071 and N00014-15-1-2792) and research proposed by several members of this team and currently under review by the Minerva project. Quantitatively, we will extend this work with a particular focus on the areas of financial interdependence, alliance dynamics, and firm-level effects. We will also draw on the strategic and regional expertise of our Air Force Academy team members to apply our work to the specific challenges that the United States faces in managing strategic competition with Russia and China. We will integrate our data-driven insights with theory and case expertise to develop a decision-guidance framework for each country, allowing decisionmakers to evaluate economic statecraft options in real time as interdependence evolves and new crises emerge.

This Project is led by Civilian Principle Investigator: Erik Gartzke, of the University of California at San Diego and the PME Principle Investigator: David Sacko of the United States Air Force Academy. Other key researchers include Co-Principal Investigators; Jack Zhang, Assistant Professor at the University of Kansas; Ben Graham, Associate Professor at the University of Southern California; Neil Narang, Associate Professor at the University of California at Santa Barbara; and Paul Bolt, Department Head, Professor of Political Science at the US Air Force Academy.

[TAI, T] Dr. Kosa Goucher-Lambert, University of California – Berkeley, and Dr. Jonathan Cagan, Carnegie Mellon University

Developing an Extendable Bi-Directional Model of Human-AI Trust for Joint Action

Our team has developed an adaptive model of human trust in collaborating AI teammates that is a dynamic entity which evolves over time and experience. We will present our proposed work under the Trusted AI Challenge Series, in which we plan to model the bi-directional trust between collaborating human and AI teammates and study the evolution of trust when members of the team have different and competing objectives. Our test scenario involves a drone design platform developed under DARPA's ATeams program, allowing heterogeneous AI/human hybrid teams to engage in the configuration design of a drone. Following the 1-year funding period we will deliver: (1) a modified problem-solving drone design platform with tunable AI agents and integrated dynamic trust models; and (2) an updated human-AI trust model extended to be bi-directional and to three agents.

[M, I, T] Dr. Jonathan Gratch and Dr. Nate Fast, University of Southern California

Title: Organizational Implications of Autonomy

Delegating tasks to autonomous machines and using them to mediate work interactions can offer many benefits, but at what cost? In human organizations, delegating tasks to others increases moral distance from the consequences of one's actions (Chugh, Banaji & Bazerman 2005). For example, company management may feel more comfortable allowing their subsidiary to engage in price gouging than if they were to set prices directly, both because the consumers are more psychologically distant when acting through intermediaries, and because consumers are more likely to assign blame to the intermediary than the parent company. We will report the results of a series of studies that illustrate this form of "moral fading" occurs when people act though autonomous agents. These findings illustrate potential risks inherent in human-machine teamwork but also highlight mechanisms that may help mitigate these risks.

[I] Dr. Julia Hirschberg, Columbia University, and Dr. Sarah Ita Levitan, CUNY

Identifying and Analyzing Right and Left-Leaning Group Videos on Social Media using Multimodal Features

We are collecting right- and left-leaning groups' videos from YouTube, Bitchute, 4Chan and Vimeo to identify which aspects of their content and presentation make these videos more popular and also potentially more persuasive. To date we have collected videos for and against Antifa and other anti-Fascist videos (14,101; 601), Black Lives Matter (13,648), Proud Boys (1431), Oath Keepers (589) and QAnon (1284) and manually labelled subsets for style, stance toward the group, persuasiveness, techniques used and other features. We have also extracted publicly-available features from a number of these videos including titles, descriptions, time of upload, captions and ASR transcripts, topic categories, and users' likes, dislikes, comments, and views. We are currently using these to automatically identify information such as the stance of the video (for or against a group), changes in popularity and in the sentiment of viewers toward the videos over time, correlating these changes with major events as well. We are also extracting text and audio features from videos and their comments to begin developing multimodal Machine Learning models for use in identifying different types of videos (e.g. pro and anti a group, extremely popular or unpopular) and potentially to use to identify new radical groups and to track their success. We will also be crowdsourcing surveys of subsets of these videos to understand how persons with different demographics and personality types perceive and are potentially influenced by different groups and different types of videos.

[T] Dr. Julia Hirschberg, Columbia University, and Dr. Sarah Ita Levitan, CUNY

Whom Do We Trust in Dialogue Systems?

It is important for computer systems today to encourage user trust: for recommender systems, knowledge-delivery systems, and dialogue systems in general. What aspects of text or speech production do humans tend to trust? It is also important for these systems to be able to identify whether in fact a user does trust them. But producing trusted speech and recognizing user trust are still challenging questions. Our work on trusted and mistrusted speech has produced some useful information about the first issue, exploring the types of lexical and acoustic-prosodic features in human speech that listeners tend to trust or to mistrust. Using the very large Columbia Cross-cultural Deception Corpus we created to detect truth vs. lie, we created a LieCatcher game to crowd-source a project on trusted vs. mistrusted speech from multiple raters listening to question responses and rating them as true or false. We present results on the types of speech raters trusted or did not trust and their reasoning behind their answers. We then describe ongoing research on the second issue: How do we determine whether a user trusts the system and do aspects of their speech reveal useful information?

[T] Dr. Leanne Hirshfield, University of Colorado (Dr. Mark Costa, Syracuse University)

Development of a Remote-fNIRS Device for use in the human performance and human-agent teaming domains

Recent advancements in biotechnology have resulted in brain measurement devices that can non-invasively measure the functioning brain in people's natural environments. Functional Near-Infrared Spectroscopy (fNIRS) is such a technique, which measures the hemoglobin signatures related to neural activation. With the potential to monitor people's mental states non-invasively and in real-time, researchers have used fNIRS devices to measure a myriad of cognitive and emotional states in operational settings. With an eye toward deployment of fNIRS devices 'in the wild', we have developed a camera-based remote-fNIRS system capable of taking fNIRS measurements from a distance of roughly ½ meter from participants. In this presentation we'll present the hardware and software underlying the device, and we'll describe research conducted to date to overcome signal quality issues introduced by participant movement, which seriously limit our ability to take measurements in naturalistic settings. Lastly, we'll give an overview of complementary human-agent teaming research in our lab, where we envision remote brain measurement to play a key role in the development of adaptive systems to support human-agent teams as they operate in real-time.

[T] Dr. Pascal Hitzler, Kansas State University (Dr. Mateen Rizki, Wright State University)

Toward Undifferentiated Cognitive Agents

In this project, we are developing a cognitive system capable of independently acquiring most required task knowledge and skill to perform a set of tasks. Our proposed approach begins with the development of a foundational and generalizable cognitive system that can be transformed into a specialized cognitive agent through written instruction, interactions with its trainers, task experience, and developer intervention (when needed). In this vein, we have developed an undifferentiated agent

(uAgent) that is a set of general-purpose computational cognitive capacities enabling it to read task instructions, iteratively interact with trainers to fill gaps in task knowledge, generate the requisite task knowledge from the instructions, and complete multitasking scenarios of varying complexity. We are working to specialize the agent to desired levels of task proficiency using the Autonomous Research System (ARES).

This project is one of a set of three tightly connected (and jointly proposed) projects. The other two are: "Toward Undifferentiated Cognitive Agents for Diverse Specializations" (Air Force Research Laboratory; PIs: Myers (RH) and Maruyama (RX)), and "Toward Undifferentiated Cognitive Agents: Translating Instructions to Knowledge" (Drexel University; PI: Salvucci). Our part of the project work focuses on the detection and resolution of gaps between instruction content and comprehension capability of the agent.

[T] Dr. Nhut Ho, California State – Northridge

Extending Theoretical Trust Modelling: A Field Study of Heterogeneous Human-Machine Teams Operating in Contexts with Real Users, Real Systems, and Real Consequences

Our study seeks to:

- Obtain foundational lessons on how trust evolves in real R3 human machine team systems (R3 stands for real users, real system, real consequences)
- Identify how technology and non-technology-related factors influence trust evolution
- Validate and extend extant theoretical trust models
- Formulate hypotheses and research questions, and identify ways to transition research results to applied research

Our study's contexts include three R3 human machine team (HMT): 1) NASA-sponsored Jet Propulsion Laboratory's (JPL) Mars Helicopter HMT, consisting of a helicopter scout, Perseverance rover, human mission controllers, scientists, and engineers; 2) DARPA-sponsored JPL's SubTerranean (SubT) HMT, consisting of a team of robots, human operator, data analysts, engineers, and human mission operators; and 3) DoD-sponsored NASA Armstrong's Traveler, consisting of a team of drones, agents (e.g., Moral Compass), and operators.

Our research methodology involves: a) applying case study methods (interviews, surveys, and participant observations) and Rapid Assessment Procedures in three interrelated case settings each with multiple participant group cases, including 1) Mars Helicopter and Traveler (i.e., operators, engineers, scientists, managers/leaders, and sponsor) and 2) SubT (i.e., operators, engineers, scientists, managers/leaders, sponsor, and regulators); and b) an iterative three-step approach to compare and contrast literature review data with emergent data across the participant group cases and three R3 HMT case settings to validate and extend a theoretical trust model.

In this year AFOSR Annual Program Review, we will present our progress in developing instruments for the case study methods, our collaboration with the JPL and NASA communities working on the three R3 HMTs, and our preliminary characterization of the heterogeneous HMT and its implications for trust studies.

[T] Dr. Colin Holbrook, University of California - Merced and Dr. Alan Wagner, Pennsylvania State University

Trust in Machine Agents Under Realistic Threat

Individuals working in dangerous situations (e.g., military personnel, police officers, firefighters) may be susceptible to threat-biased assessments of the machine agents with which they increasingly team. Threat biases favoring conformity may facilitate coordination when confronting threats by promoting integration of machine recommendations, yet may erode decision-making in some contexts by potentiating over-reliance. Our research objective is to gauge the extent to which anthropomorphic physical and behavioral characteristics determine trust in the recommendations of robots and software agents under life-or-death circumstances. As the COVID-19 pandemic rendered planned laboratory research impossible, we adjusted our strategy in Year 1 to focus on online studies designed to inform our future lab-based human-robot interaction studies as well as to leverage the affordances of internet-based research to explore trust in screen-mediated agents. In this presentation, we will briefly review the results of three experiments manipulating the apparent gender of anthropomorphic robots, two studies manipulating apparent white versus Black race, and two studies assessing the role of anthropomorphism and threat on trust in a dyadic decision task. Summarizing key findings which proved replicable, we found that i) anthropomorphic robots and humans were perceived as comparably competent and influential in a behavioral team decision-making paradigm, ii) non-anthropomorphic robots were perceived as substantially less trustworthy than anthropomorphic robots, iii) manipulations of the apparent gender or race of anthropomorphic robots exhibited negligible effects on trust, and iv) appraisals of robots as intelligent predicted trust with regard to domains such as care of information or objects, whereas perceptions of robots as warm and alive predicted trust in their ability to provide care for living agents (e.g., pets, children). These results will be discussed as they guide our lab-based research designs moving forward. In addition, we will share progress made in the development of novel online testbeds for manipulating effects of threat-salience and anthropomorphism on trust in machine agent recommendations with regard to navigation, threat-detection, and the use of force. Finally, laying the foundation for the face-valid, realistic VR paradigms to be employed in Years 2 and 3, we will share virtual environments and robot avatars acquired, modified or created by our team to date.

[M, T] Dr. Michael Horowitz, University of Pennsylvania

The Disruptive Effects of Autonomy: Ethics, Trust, and Organizational Decision-making (MINERVA)

Rapid advances in autonomous systems raise fascinating behavioral questions. In particular, for the Department of Defense, understanding how factors like trust, risk, and organizational incentives could shape the development, use, and effectiveness of autonomous systems will be critical. In order to answer these questions, experts will need to move beyond the purely technical factors and models of effectiveness traditionally used in military analysis and utilize a broader range of behavioral science tools and organizational theory. These questions have great relevance for the Department of Defense

due to existing investments in autonomous systems and human-machine teaming, the broad framework of the 3rd Offset, and growing knowledge about significant investments in autonomous systems by China, Russia, and other countries. This project seeks to understand the human, organizational, and political factors that will affect the willingness of individuals and bureaucracies to adopt autonomous systems, and the potential consequences of these attitudes. How do knowledge and familiarity, as well as the consequences for job prospects, influence attitudes towards autonomous systems? What characteristics will guide when U.S. military personnel and decision-makers trust autonomous systems, and what are the implications of any trust gap relative to other types of systems? How would different types of autonomous systems affect the willingness of the U.S. public and decision-makers to support the use of force? Finally, how will the use of autonomous systems influence crisis bargaining and coercion, particularly in flashpoints such as the South China Sea? To answer these questions, this project will bring together theories on adoption capacity and military technologies (Horowitz 2010), theories on the societal effects of emerging technologies (Juma 2016), and theories on what drives public attitudes concerning emerging technologies (Schneider and Macdonald 2016, Horowitz et al. 2017), to build a broader understanding of the disruptive behavioral effects of autonomy.

[T, I] Dr. Michael Horowitz, University of Pennsylvania

Small Steps and Giant Leaps: Remoteness, Bias, and Decision-making in Space

What cognitive biases affect military operations and decision-making in the space domain?
Recent research on emerging technologies has shown that automated and autonomous systems can warp human decision-making in predictable ways, including automation bias, which can have serious consequences for defense. As use of space grows and space becomes a focal point for competition, it is critical to understand how questions of trust, confidence, and biases will influence the adoption and use of different space strategies. Extreme distance, remote control, degrees of automation, and even cultural understandings shaped by science-fiction literature and films could all distort leaders' and operators' perceptions and decision-making capabilities in space. These factors increase uncertainty about the space domain, which generates friction. We focus in particular on three ways that space is remote that could shape strategy and behavior: physical distance, difficulty of access, and psychological unfamiliarity. This proposal unpacks these questions through a multimethod approach that includes survey experiments, wargames, quantitative text analysis, historical analysis of the rise of new domains of warfare, and case studies of cross-national developments of space capabilities and perceptions of space as a domain. The results will have critical relevance for questions of space strategy and operational planning. Additionally, the results could shed light on broader questions surrounding norms of behavior in space, how to handle questions of the "commons" surrounding space debris, and space exploration.

[I, X] Dr. Bailu Jin and Dr. Weisi Guo, Cranfield University

Networked Social Influence and Acceptance in a New Age of Crises

In an age of emerging disruptive crises (COVID) and technologies, understanding influence in traditional and online societies is critical to defence and technologies. We know that influencers play a vital role in influencing the opinion of others and creating social acceptance. This is a vital mechanism in the dissemination of new ideas. However, what we do not understand is the effective combined

mechanisms that stem from how the influencer behaves and where he/she stands in a multi-scale social network. The research aim is to understand the joint impact of individual behaviour and the wider influence from social network structure. Traditional social influence can be gauged by complex network analysis, where eigenvectors point towards influential locations on the social network. Whilst this has led to useful metrics such as PageRank, it neglects the complex nonlinear dynamic interactions between people and the multiscale topological factors that contribute towards a more sophisticated picture. Here, we set out the ambition to create a modeling framework, grounded in real and diverse social data, on how to combine complex social network analysis with nonlinear human behaviour dynamics. This has the advantage of being able to uncover hidden influencers. This has widespread applications in counterterrorism, social media analysis, and understanding how we can effectively spread policies and create acceptance to technologies in an age of emerging crises.

The objectives are: 1. Who are the real influencers: finding influencers on social networks as a function of network topology and nonlinear influence behaviour dynamics. 2. Explainability using AI: identify what parameters cause influencers to be influential and personal behaviour dynamics.

This has the advantage of being able to uncover hidden influencers. This has widespread applications in counterterrorism, social media analysis, and understanding how we can effectively spread policies and create acceptance to technologies in an age of emerging crises (e.g. COVID).

[M, I] Dr. Neil F. Johnson, George Washington University

Total War: Multi-Agent Network Theory of Connective Action in a Cross-Domain Coupled World

This talk will present progress so far in Year 1 of this Minerva project, with the additional goal of developing new synergies with other projects within the trust and influence program. Overall, this Minerva project aims to develop the fundamental science of how trust and influence develops in the cross-domain, cross-scale, hybrid online-offline world 'left of boom' of Total War. Grounded in social science, it will develop a quantitative, data-driven description for cross-domain and cross-scale tipping points -- like coupled forest fires. Potential benefits to DoD include a deeper understanding of how to measure and model bottom-up and top-down malign influence, as well as peer and near-peer competition. It will also help better understand, measure and model the stability of, and challenges to, national and global security. In this way, the project will achieve the goals of helping expand understanding of peer and near-peer competition and foreign malign influence, as well as increasing increase understanding of the social aspects that underlie security and stability and how they may result in challenges to national security. By so doing, it will attempt to make foundational contributions to basic social science in social media analytics for foreign malign influence and the interactions of peer and near-peer state actors.

Specific application areas to be discussed in this talk, include an analysis of online hate and extremism. The talk will also discuss 'soft power' from COVID vaccines and future boosters, as well as the 'anti' and 'hearts and minds' subpopulations which are combining to create a new form of multi-sided arms race.

[M] Shinobu Kitayama, University of Michigan, and Michele J. Gelfand, Stanford University

Complacency under Germ Threat: Interdependent Self-Construal Moderates Neural Responses to Norm Violations

It is both remarkable and deplorable that, in the current COVID-19 pandemic, many individuals are so highly complacent, resulting in avoidable casualties. It is therefore urgent to explore factors that foster this complacency. Here, we hypothesized that social embeddedness gives a sense of security. We thus anticipated that the construal of the self as interdependent and thus socially embedded would reduce the psychological impact of an impinging germ threat. Participants who varied in interdependent self-construal were either primed or not with a germ threat. They then judged how norm-violating various behaviors were, with their electroencephalogram (EEG) monitored. The behaviors were either norm-violating or normal. We tested two parameters of EEG indexing (i) the detection of norm violations (N400, known to increase by semantic incongruity) and (ii) vigilance to norm violations (alpha power, known to increase by external attention). In the control priming condition, we found that interdependent self-construal predicted increased neural reactions to norm violations. However, in the germ priming condition, interdependent self-construal predicted reduced neural responses to norm violations. Self-report of norm-violation severity showed no such effects. Our results indicate that social embeddedness is a double-edged sword. It can protect people against various threats. However, precisely because it is sometimes protective, it can induce complacency, producing an unwarranted sense of security. In addition to identifying a condition conducive to complacency under threat, our work has provided a neural instrument to assess people's sensitivity to norm violations in the absence of any visible effects in self-report.

[T, DR] Dr. Zhaodan Kong, University of California – Davis, and Dr. Allison Anderson, University of Colorado

Network-based Neurophysiological and Psychophysiological Metrics of Human Trust Dynamics When Teamed with Autonomy

In this talk, we will briefly present our recent research efforts to develop neurophysiological and psychophysiological measurement based trust metrics with a particular focus on dynamic assessment of trust in human-autonomy teaming tasks. Our project seeks to breach key scientific and technological barriers that hinder machines from effectively and dynamically inferring their human partner's trust. Specifically, this project integrates the fields of cognitive science, network science, and human factors to (1) develop an experimental methodology to study human trust dynamics in human-autonomy teaming scenarios and (2) investigate if a series of network-based metrics that are derived from neurophysiological, e.g., EEG and fNIRS, and psychophysiological measurements, e.g., heartrate variability and skin conductance, can be used to predict human trust dynamically. This project is partially funded by a DURIP grant from AFOSR and a Space Technology Research grant from NASA.

[M, I] Dr. Gizem Korkmaz, Associate Research Professor, Social and Decision Analytics Division, Biocomplexity Institute, University of Virginia and Dr. Read Montague, Professor and Director, Human Neuroimaging Lab, Fralin Biomedical Research Institute, Virginia Tech

The Dynamics of Common Knowledge on Social Networks: An Experimental Approach

We study protest as a collective action problem where an individual wants to participate only if joined by “enough" like-minded others. In game-theoretic contexts, coordination requires that agents know about each other’s willingness to participate and that this information is common knowledge among a sufficient number of people. Stylized models combine social structure and individual incentives together, and provide a rigorous game-theoretic formalization of common knowledge, and the characterizing network structures. These models emphasize simple node-to-node or bilateral communication, or study the effects of “richer" online communication mechanisms, such as Facebook or Twitter. To date, very little is known about how well these models explain phenomena in practice or how individual and behavioral factors influence group engagement. Our work aims to empirically test novel hypotheses at both individual and group levels by conducting controlled human subjects experiments in three different environments (laboratory, online, and neuroimaging experiments) that inform models of collective behavior. The objectives of this integrated framework are (i) to characterize how different communication mechanisms can facilitate actionable common knowledge through local interactions, (ii) to understand the effect of network structure, and (iii) to quantify individual and group level behaviors, and neural processes that affect its formation. The results of our online experiments demonstrate statistically significant results by network structure and communication type with clique structures and wall communication leading to higher levels of group participation. Our neuroimaging efforts have focused on two different experiments. The first neuroimaging experiment is a scanner version of the behavioral task discussed above. Rather than displaying the multiple pieces of information required to make a decision all at once they are displayed sequentially so that the neural responses to the separate pieces of information can be teased part. We are specifically interested the differences in neural responses across multiple information conditions (local vs. global information; clique vs. circle topologies on the graph). In the second experiment, a participant undergoing fMRI plays a graph coordination game (a maximum matching task). The other participants on the nodes are computer bots. In a round of the task the participants and bots each send an invitation to one of their neighbors. The goal is to maximize the number of matches in the network. Here we are interested in the neural representations of learning local configurations in the context of maximizing the global score. We continue to collect the neural data and have begun to analyze it. Together, our results will provide insights on how network structure and communication mechanisms shape collective action.

[M, I] Anthony F. Lemieux, Georgia State University
Carol Winkler, co-PI (Georgia State University); Jonathan Pieslak, co-PI (City College of New York); Akil Awan, co-PI (University of London); Weeda Mehran, co-PI (Exeter University, previous Post-Doctoral fellow at GSU); Ben Miller (Emory University); Nelly Lahoud (Consultant); Jarret Brachman (Consultant); Humera Khan (Consultant).

Mobilizing Media: A Deep and Comparative Analysis of Magazines, Music, and Videos in the Context of Terrorism

Mobilizing Media utilizes three approaches to analyze the media campaigns of predominantly MENA-based extremist groups. The first examines audience cultivation through identification of core themes and persuasive elements, their level and nature of use by different groups, and the events that prompt them to change over time. The second examines the unique information conveyed by textual, visual, aural, and sonic elements of the media campaigns, as well as strategies of interactive reinforcement the groups deploy across the different modalities. The third examines how the groups' use of information, motivational appeals, and behavioral skills maximizes their chances of producing behavioral change. This presentation will summarize key findings over the life of the project to date, and will outline future directions in the final year of the project.

[I] Dr. PI: Kristina Lerman, University of Southern California

Detecting and Mitigating Adversarial Influence Operations in Networks

Successful response to societal disruptions requires sustained behavioral change. However, the divergent responses to the Covid-19 pandemic in the US showed that political polarization and mistrust of science can reduce public's willingness to adopt mitigation measures, such as mask wearing and vaccination. To better understand this phenomenon, we use a large dataset of tweets related to the Covid-19 pandemic to quantify the partisanship of social media users and their propensity for sharing anti-science misinformation. We find that these dimensions are correlated, with conservatives more likely to disseminate misinformation. We study users' exposure to information, which is shaped by their social networks and the algorithms social media platforms use to recommend relevant posts. Our analysis reveals multi-dimensional echo-chambers. In general, users surround themselves with people with similar views; however, users with hardline political views are more likely to spread misinformation, regardless of the veracity of their exposure. In addition, I describe a novel change detection algorithm that automatically discovers sudden changes in the topics of online conversations, which malicious actors can exploit to amplify polarization. This research aims to increase resilience of online information ecosystems to societal disruptions and malicious manipulation.

[I] Dr. Yu-Ru Lin, University of Pittsburgh, and Dr. Lexing Xie, Australia National University

Linking Online Attention to Measurable Actions: Linking Political Attention on Twitter and YouTube

We present a measurement study that links collective attention across two social media platforms -- Twitter and YouTube, focusing on videos on controversial political topics. We develop data collection procedures that link Twitter posts and topical YouTube videos, yielding three new cross-platform datasets covering the controversial topics Abortion, Gun control, and Black Lives Matter over 16 months. We propose a set of video-centric metrics across both platforms to qualitatively compare the collective online attention from different political groups. We observe that among the most popular videos, left-leaning ones are overall more viewed, more engaging, but less tweeted than right-leaning ones. We find that attention unfolds quickly on left-leaning videos, but spans a longer period of time for right-leaning videos. We also correlate online activities with a collection of offline signals.

[I] Dr. Yu-Ru Lin, University of Pittsburgh, and Dr. Lexing Xie, Australia National University

Linking Online Attention to Pandemic Responses: Media Engagement and Misinformation during the Pandemic

Within the US, we measure daily, state-level aggregated misinformation engagement from over 242 million tweets in the United States between early March and late June 2020, and compare how people's movement changes across different types of locations (e.g., recreation, grocery, and workplaces), non-pharmaceutical interventions, COVID-19 cases and deaths, along with population demographics and partisanship. Our preliminary causal analyses show that a higher level of online COVID misinformation engagement has led to a lower extent of mobility reduction, controlling for confounders and spillover effects, suggesting the role of online misinformation in altering people's adherence to pandemic controls. Internationally, we leverage a unique high-volume longitudinal twitter dataset to measure the geographical engagement of media outlets across the political (and factual) spectrum across 6 continents. These measurements quantify the reach of different media outlets, and allows one to quantify twitter user engagement internationally on the US left-to-right spectrum. We intend for the new understandings of the propagation of misinformation, and the profile of global media engagement to contribute to our collective battle to curtail the global pandemic.

[T] Dr. Nathan McNeese, Clemson University

The Spread of Trust and Distrust in Distributed Human-Autonomy Teaming Constellations

This presentation will outline upcoming work focused on understanding how trust and distrust spread in a multi-human-autonomy team (HAT) constellation. To prepare for the coming expansion and integration of human-autonomy teamwork, this research establishes a solid foundation to investigate the ability of trust to travel between teammates and teams in environments consisting of multiple HATs interacting. The work consists of: 1) repeated, mixed-methods empirical experiments that observe trust both within and between HATs; 2) an emphasis on ensuring HAT research is conducted in applied environments that are representative of those seen in DoD initiatives; 3) the iterative development and validation of quantitative measure specifically designed to capture and quantify the spread of trust between teammates within and across different teams; and 4) the landmark application of trust repair techniques in a multi-HAT constellation with the goal of preventing the organizational-wide spread of distrust for autonomous teammates.

[T] Dr. Nathan McNeese, Clemson University

Considerations of Ethical and Unethical Behavior on Trust in Human-Autonomy Teaming

As human-autonomy teaming continues to be conceptualized and grow in relevance, the role of ethical and unethical behaviors impact on trust must be considered. This presentation highlights work focused on: 1) an experiment that seeks to understand how ethical and/or unethical autonomous teammate agents' behavior directly impacts the dynamic behavior of human trust in a human-autonomy teaming synthetic task environment, and the effectiveness of trust repair strategies in this domain; 2) focus group data to understand why ethicality and trust repair may or may not have an effect on human behavior and outcomes; 3) a factorial survey to determine how the type of unethical behavior influences trust; 4) interviews with subject matter experts (Air Force pilots) to discover their insights on trust,

ethics, and trust repair with a hypothetical autonomous teammate. Stemming from each of these research studies, a myriad of results will be presented that help to better understand the relationship of ethics and trust in human-autonomy teaming.

[I] Dr. Eduardo L. Pasiliao, AFRL/RWWN

Modeling and Learning Highly Nonlinear Interactions within Social Media Networks

Social media networks have become a truly planetary-scale phenomenon nowadays. Interactions among social media users (nodes in social media networks) are complex and multi-faceted, with many known and unknown factors that may need to be taken into account. Although there are known mathematical models for social network interactions (i.e., the classical linear threshold models), interaction and influence dynamics in social media networks in reality occur in a highly nonlinear (non-monotonic) fashion. This is because not only the immediate neighbors of an individual user may affect "influence" (or "activation") of a user, but also other factors, such as potential influence of more distant (multi-hop) neighbors of a node, clustering effects, as well as global topology of a social network, need to be taken into account from mathematical modeling and computational perspectives. In addition, since social media networks are not "isolated" systems but exist in close interactions with the physical environment, external factors (e.g., geo-spatial aspects) of social media interactions also need to be incorporated into the considered models. Another interesting aspect in this context is related to human cognitive limits in terms of accepting and maintaining interactions in social media network settings (e.g., although some people have a large number of online "friends", our preliminary findings suggest that there may be a cognitive limit on the number of "meaningful online friendships", similar to Dunbar's number hypothesized for "real-life" social interactions). Overall, our approach is to enhance the models based mostly on 0-1 adjacency matrices, which are often used in social network models, with more complex functional relationships between nodes, groups, and clusters (e.g., , Graph Laplacian, neighborhood topology, strong and weak ties), reflecting heterogeneous factors that may affect interactions within social media networks. Thus, the main goal of the proposed project is to use and synthesize mathematical modeling (e.g., graph theory, optimization) and machine learning (e.g., artificial neural networks) techniques to comprehensively address various factors driving social media dynamics and develop models and algorithms that would reflect social media interaction processes more realistically than previously developed techniques.

[T] Dr. Elizabeth Phillips, George Mason University, and Dr. Bertram F. Malle, Brown University

Moral Justification to Foster Human-Machine Trust

Robots and other autonomous agents are increasingly being used in domains such as space operations, cyber defense, disaster response, and medical care, and are envisioned to directly collaborate with human partners in ways that resemble human-human teams. But, all human communities, groups, and teams have norms that influence and regulate behavior, so autonomous agents that join these communities must be responsive to norms as well—they must know and follow the norms of their community. But even if we succeed in giving autonomous agents such norm competence, we are faced with a significant challenge: Norms can conflict with each other. Often, the only way to resolve conflicts between norms is by deciding to uphold one norm—the more important one—and to violate the other, less important norm. This means that whenever an agent (human or machine) resolves a norm conflict,

it must commit a norm violation. People respond to such violations with moral disapproval and loss of trust. In this project we investigate one powerful tool humans use—and autonomous agents should use—to mitigate such moral disapproval and repair lost trust: justifications. When an agent must violate a norm in order to resolve a norm conflict, a justification explains why the agent did act in this way and why anybody that shares the community's norms should act in this way. In a series of experiments, we will demonstrate that, after resolving a norm conflict and committing a norm violation, an autonomous agent that justifies its actions—similar to a human who does so—will reduce the moral disapproval and repair the loss of trust that normally results from norm violations.

[I] Dr. Chris Martens, North Carolina State University

Towards Abstractions for Interactive Social Simulation

Simulating the decision-making behavior of diverse groups of humans within a shared environment has long been a grand challenge of artificial intelligence, as well as of great interest for game design and social science research. However, recent advances in data-driven machine intelligence have not brought much new to bear on this set of important challenges, in part because of the diversity of contexts -- for example, the traffic patterns of people commuting to work is a very different context (with correspondingly different prediction problems) from how people formulate and revise opinions of one another through conversation. In this talk, we present an analysis and taxonomy of social simulation systems, and outline our ongoing work on distilling a set of reusable computational abstractions for social behaviors.

[T] Dr. Christopher Myers, 711 Human Performance Wing (711th HPW) (AFRL/RH), and Dr. Benji Maruyama, Materials and Manufacturing Directorate (AFRL/RX)

Toward Undifferentiated Cognitive Agents

In this collaborative project, we are developing a cognitive system capable of independently acquiring most required task knowledge and skill to perform a set of tasks. Our proposed approach begins with the development of a foundational and generalizable cognitive system that can be transformed into a specialized cognitive agent through written instruction, interactions with its trainers, task experience, and developer intervention (when needed). In this vein, we have developed an undifferentiated agent (uAgent) that is a set of general-purpose computational cognitive capacities with the ability to read task instructions, iteratively interact with trainers to fill gaps in task knowledge, generate the requisite task knowledge from instructions, and complete multitasking scenarios of varying complexity. The Air Force Research Laboratory portion of this collaboration, through interaction between the Airman Systems and Material and Manufacturing Directorates, is to leverage controlled languages for translating English instructions into machine-parsable representations, pulling the components of the system into a unified whole and evaluating development time differences between the current standard method to model development and training a uAgent across tasks of increasing complexity.

[I] Dr. Robert Pape, University of Chicago

Assessing the Impact of Right Wing Extremist Organization Propaganda on Radicalization and Support for Domestic Political Violence among Individuals with Military Service

This study investigates the factors associated with the appeal of right-wing extremist organizations (REOs) and domestic political violence among Americans with military experience. Despite vast attention paid to international terrorist recruitment propaganda, there is a gap in empirical research on how and why online videos inspire and mobilize new individuals to domestic terrorism. To understand the role of REO messaging and the unique contribution of military experience and risk factors for radicalization, we will conduct a survey study. The study will field the survey through the National Opinion Research Center at the University of Chicago on a representative sample of 800 American adults with military experience (active-duty members and veterans) under the age of 61 and a matched sample of 800 American adults without military experience (total N=1,600). The survey has two components. First, it will determine the baseline support for REOs and the distribution of common risk factors for radicalization such as ideological and behavioral predispositions, mission stress and trauma during military service; reintegration stress after military service; and social networks of close friends. Second, the survey will use an experiment to measure the impact of different video appeals to join REOs and/or participate in domestic political violence in the population with military experience using non-military population as a control. This project will identify risk factors and vulnerable sub-populations among individuals with military service and provide valuable diagnosis of the relationship between US military service and militant group recruitment, which will be useful for DoD messaging, counter-messaging, and other policies aimed at countering violent extremism.

[M, I] Dr. Robert Pape, University of Chicago

The Social and Neurological Construction of Martyrdom

The overarching goal of this project is to develop the first comprehensive social and neurological understanding of how extremist organizations propagate their messages and which audiences are most susceptible. To do this, we propose a three-year program of basic research, with an optional twoyear extension, to improve our understanding of how violent extremist organizations (VEOs) such as the Islamic State of Iraq and Syria (ISIS) construct cultures of martyrdom that mobilize support for their activities, especially their use of suicide attacks. Though scholars agree that establishing a culture of martyrdom is essential to a group's ability to maintain support for its most violent actions, we do not yet understand the social processes and conditions under which they succeed or the degree to which psychological and cultural factors predispose certain audiences to accept cultures of martyrdom

[M, D] Will Reno, Northwestern University, and Dr. Ryan Burke, US Air Force Academy

Foreign Military Training (FMT): Building Effective Armed Forces in Weak States

Our research design captures the entirety of the FMT process from design to provision to absorption stages in "weak states" (i.e., states with organizationally fragmented armed forces and militias, and

governments that often lack the political will to sustain strong armed forces), with a focus on Iraq, Mali, Niger and Afghanistan (possibly to be replaced or supplemented with Ukraine). Our research aim is to capture and explain the few intended and many intended outcomes of FMT. Insights from that research are used to explain cross-national variations, particularly in contexts where governance is shaped by patron-client relations that undercut formal institutions and procedures, and to explain in-country variations. Preliminary findings indicate FMT can produce highly effective military units and enclaves of competence in wider political contexts dominated by patronage networks. The outcome of this research aims to explain in-country variations in FMT outcomes, and to generate Maturity Models that will improve FMT training manuals and improve core courses of USAFA's Military & Strategic Studies department. This research has direct relevance to DoD missions to develop allied and partner military capacities for self-defense and coalition operations.

This project involves considerable field research among designers, providers, and beneficiaries of FMT across US, NATO partners and recipient states. Final Human Research Protection official (HRPO) Review and Component Level Administrative Review (CLAR), and Command acknowledgements were completed on 17 May 2021. These reviews are necessary conditions for human subject research. During the review process it became apparent that the Afghanistan component of this research would not include field visits.

[T] Dr. Laurel Riek, Computer Science and Engineering, University of California - San Diego

Trust Affordances in Human-Robot Teaming

When engaging in human-robot teaming (HRT) in dynamic, uncertain, environments, it is crucial that there is mutual understanding and well-calibrated trust between humans and machines. This talk will discuss recent work from my lab which explores the algorithmic side of this problem, particularly with regards to robots that can sense, understand, and make decisions under uncertainty to support HRT in critical environments. I will also discuss our recent efforts applying this basic research to building and deploying new shared autonomy systems in emergency medicine, to support improved teaming and safety during the pandemic.

[I, Y] Dr. Daniel Romero, University of Michigan

Assessing the Impact of Exogenous Shocks on User Behavior and Information Diffusion in Social Media

I will discuss an agent-based model of information diffusion that compares the impact of network structure and social identity on the spread of information. Comparing this model against a dataset of 76 lexical innovations on Twitter, we find that network topology and social identity are jointly sufficient, but not individually sufficient, to explain many key properties of linguistic diffusion.

Moreover, we find that key linguistic regions, and the pathways through which words diffuse to produce those regions, "emerge through reinforcement." That is, these pathways are relatively consistent in versions of the model where only network topology or social identity mediates diffusion, but change dramatically, aligning with known dialect regions, when the variables are taken together.

[T, Y] Dr. Dorsa Sadigh, Stanford University

The Role of Conventions in Adaptive Human-AI Interaction

Today I will be talking about the role of conventions in human-AI collaboration. Conventions are norms/equilibria we build through repeated interactions with each other. The idea of conventions has been well-studied in linguistics. We will start the talk by discussing the notion of linguistic conventions, and how we can build AI agents that can effectively build these conventions. We then extend the idea of linguistic conventions to conventions through actions. Finally, we go over a modular approach to separate partner-specific conventions and rule-dependent representations and demonstrate the effectiveness of this modular approach in a number of multi-agent and human-AI collaborative games.

[T] Dr. Dario Salvucci, Drexel University

Toward Undifferentiated Cognitive Agents: Translating Instructions to Knowledge

In this collaborative project, we are developing a cognitive system capable of independently acquiring required task knowledge and skill to perform a set of tasks. Our proposed approach centers on the development of a foundational and generalizable cognitive system that can be transformed into a specialized cognitive agent through written instruction, interactions with its trainers, task experience, and developer intervention (when needed). In this vein, we have developed an undifferentiated agent (uAgent) that includes a set of general-purpose computational cognitive capacities with the ability to read task instructions, iteratively interact with trainers to fill gaps in task knowledge, generate the requisite task knowledge from instructions, and complete multitasking scenarios of varying complexity. The Drexel portion of this collaboration focuses on translating instructions to knowledge, using a newly developed "cognitive-code" architecture, Think, as the theoretical and computational foundation for the work. Our most recent developments have seen advances in the flexibility and robustness of the instruction-translation process, as well as new explorations into applying the core ideas of the uAgent to developing novel teachable agents.

[TAI, T] Dr. Eugene Santos, Dartmouth

Quantifiable Expectability and Measurable Intentions

Human-AI calibrated trust needs quantifiable short and long term computable expectability. We define expectability as being anticipatory, understandable, and precautionary with regards to one entity's behavior by another entity necessary for joint action. Calibrated trust requires inferring of intent of and by each partner to ascertain how close or far the inferred intent is from the actual intent and hence trust. Inferred intentions in this effort involves learning the rewards and reward structures employed by the target entity as part of shared context. Deviations from the expected rewards and structures from the actual forms the computational basis for trust calibration. In this presentation we briefly present our goals and approach in this upcoming project to define, develop, and demonstrate measurable intentions for computational calibrated trust in human-AI partnerships applied to joint activity and/or shared context for the domains of intelligence and digital data analysis and tactical and strategic decision making in computer games and simulations. Our solution is founded upon new concepts and algorithms for individual (human or AI) and team-based multi-agent inverse reinforcement learning (IRL), a novel Preferential Trajectory IRL (PT-IRL) that distinguishes between multiple different individuals and teams by their behavior overcoming existing linearity relationship assumptions in IRL, and a formal definition for

quantifiable interference and intentionality gap that reveals and explains how each team member is adapting and learning during joint activity and decision-making.

[T] Dr. Nadine Sarter, University of Michigan
Graduate Students: Kevin Lieberman and Karanvir Panesar

Supporting Trust Calibration and Attention Management in Human-Machine Teams

Trust plays a critical role in determining the behavior and safety of human-machine teams. This project aims to develop a better understanding of how trust relates to attention management – these two phenomena are closely connected but have been examined separately in most studies to date. Specifically, we explore how trust and attention management co-evolve unaided over prolonged periods of time and how it can be shaped through design and training in the interest of improved joint system performance. As first steps in this project, we have reviewed the theoretical and operational literature on multi-UAV control, visited military sites and conducted interviews with subject matter experts (SMEs). Next, we developed an enhanced multi-UAV control simulation that imposes multitasking demands and presents study participants with both nominal and off-nominal scenarios. The simulator was interfaced with an eye tracker so that visual attention allocation can be traced and used to infer variations in trust levels. Using this simulator, we conducted a longitudinal study that assessed how different types of operator training and operational experience over a 4-week period may support trust calibration, attention management, and performance in human-machine teams. Due to the COVID-19 pandemic, completion of the data collection for this experiment was delayed. As a result, at last year's review, we presented partial, preliminary findings only; this year, we will discuss results from the full dataset. In addition, we will report findings from a more recent study that assesses how the presentation of confidence information associated with a UAV's automated target identification performance impacts operator trust and attention management. In particular, this research assesses the impact of a machine's framing of its own estimated reliability as "confidence" or "uncertainty", the effectiveness of various representations of confidence and uncertainty information, and the effect of temporarily changing the accuracy of a machine's estimated confidence and uncertainty information (i.e., when there is a mismatch between the machine's estimated and its true reliability).

[T] Dr. Matthias Scheutz, Tufts University

Enabling Trusted Human-Like Artificial Teammates

The aim of this project is to lay the foundations for future autonomous systems in mixed-initiative human-machine teams to be able to operate at human-like levels of interactivity and effectiveness. We are integrating measurements of neurophysiological signals from human teammates such as fNIRS and EEG with multiple additional measurements (physiological, linguistic, behavioral) and situational contextual information to classify various individual and team cognitive states.

Classified cognitive states of all teammates are individually tracked and fused in real-time, and integrated into a shared mental model which uses advanced probabilistic "theory of mind" representations to capture team and task states with their associated uncertainties, and supports the decision-making and behavior adaptations of the autonomous artificial teammates. In this talk, we will report on our progress in all of the above areas, in particular, on the challenges of classifying cognitive states based on physiological signals and context.

[T, DR] Professor Thomas Schnell, Operator Performance Laboratory (OPL), Iowa Technology Institute (ITI)

Intelligent Tactical Autopilot to Advance Basic Understanding of Human Reliance and Teaming in Mixed Human-Machine Teams

This DURIP seeks to purchase an autopilot system with an open control interface and architecture for integration into the L-29 Flight Test Aircraft of the University of Iowa Operator Performance Laboratory (OPL). The autopilot kit is made by Collins Aerospace and comprises of the following components: 1). SVO-5000 Smart Servos for elevator, aileron, rudder, and throttle, 2). REU-6000 High Integrity Autopilot Computer and Shape Monitor, 3). AHC-4000 Attitude and Heading Reference, 4). ADC-3020 Air Data Computer, and 5). GPS-4000S.This configuration will collectively be referred to as the Intelligent Tactical Autopilot (ITAP). The purpose of the ITAP is to affect the legacy flight control linkages of the OPL L-29 research aircraft through an open interface that can be connected to a General Purpose Processor (computer). This interface enables the legacy airframe to perform autonomous flight modes for use in flight test research of mixed human-automation teams. The envisioned flight tests will always be conducted with a rated pilot onboard, acting in the role of a Safety Pilot (SP). To ensure straightforward airworthiness certification, the DURIP ITAP components and software are flight grade certified avionics and are configured to ensure that the system is fail-passive and safe for the onboard pilot(s). A close variant of this configuration was provided by Collins to Aurora Flight Sciences for integration in the AACUS UH-1 aircraft. The DURIP ITAP has dual lane redundancy, features provisions for configurable aircraft envelope protection, and has a dual-redundant shape monitor that ensures that dangerous or invalid commands cannot be executed by the servos. The DURIP ITAP features numerous methods for rapid disconnection through software guards, pilot disconnect switches, electrical clutches, power disconnect, and shear-pins. Conventional autopilots are closed systems that cannot typically be interfaced to research control algorithms. Those autopilots have built-in modes to perform gentle maneuvers commonly associated with instrument flight procedures. This limits their utility for the type of research needed under the trust and influence pillar. The DURIP ITAP is specifically designed with an open command interface and approaches full envelope authority in all axes. The DURIP ITAP will expand the already significant research instrumentation of our L-29 flight test aircraft and will make it the first such airborne testbed to support ongoing and future DoD research projects. Currently, there are no airborne testbeds available with an open architecture autopilot such as the DURIP ITAP and which have comprehensive integrated human-behavioral research equipment. The OPL L-29 aircraft are already equipped with a comprehensive suite of physiological and flight state monitoring equipment to measure behavioral metrics of the aircrew during tactically relevant maneuvers. These aircraft also have an F-35 Helmet Mounted Display (HMD) and a large format Head-Down Display (HDD) touch screen to study and analyze aircrew behavior throughout dynamic maneuvering envelopes in a real-world airborne context. OPL's aircraft also have simulated weapons models that are adjucated in live flight and which can be employed by the human aircrew and/or autonomous agents. What the aircraft are currently lacking is an autopilot with an open interface that can be used to close-the-loop with on and off board control agents. Adding this autopilot to our already heavily instrumented L-29 flight test aircraft will provide unique and cost-effective assets that can be employed by DoD, industry, and academia to conduct this foundational behavioral research in a realistic and safe manner. The flight test aircraft are currently used

by various graduate research programs at the University of Iowa, AFIT, and the USAF Test Pilot School. The acquisition of this autopilot will further enhance the University of Iowa's ability to educate students through research in disciplines important to DoD missions.

[M, I] Dr. Shade Shutters, Arizona State University

Growing Chinese Economic Power and the Exacerbating Effects of U.S. Economic Interdependence

The foundation of U.S. power is its economy and that economy is critically dependent on a stable international order. Yet, globalization of trade and technological advancement over the past two decades have created an international economic system that poses several significant and interrelated threats to U.S. national security. First, the interdependent nature of this system is a source, not only of strength, but also of vulnerability. The global nature of today's economic systems means a disruption in one place can swiftly cascade across the world, disrupting global supply chains and national economies. Second, global economic interdependence has virtually replaced linear chains of cause and effect with a complex network of causal impacts. U.S. national policies such as tariffs, which once had largely predictable outcomes, now often lead to unanticipated and undesirable consequences for the U.S. economy. Third, the current structure of global economic systems exposes the U.S. to potential acts of coercion and extortion by other countries, including key trading partners. China, in particular, is increasingly adept at using such economic weapons. However, in the peer-reviewed academic literature there exist two competing views of how interdependence affects the vulnerability and resilience of social systems. While one school highlights the benefits of interdependence, and even promotes economic interdependence as a path to global peace, a second school focuses on the dangers of deep interdependence. This lack of consensus has persisted for years with little progress towards generalizable theories. If knowledge in this area is to progress these divergent viewpoints must be reconciled. This project addresses this intellectual discord and associated security needs by examining how economic interdependence induces vulnerability and creates threats to national security. We will construct (1) quantified metrics, models, and indices of global economic system and industrial sectoral vulnerability, and (2) a multilayer network analysis of U.S. interdependence with China and how those networks combine to exacerbate system vulnerabilities.

[T, X] Dr. Sónia Sousa, Tallinna Ulikool

Factors Influencing Trust in Technology

GRANT13361881 SONIA SOUSA DATE: 1.09.2021	TrustedID

The main goal of the TrustedID project is investigate cross-cultural differences in users' trust and, based on it, propose a framework to assist practitioners in the development of trustworthy systems. The first aim (1) is to explore through the lens of the Human-Computer Trust scale (Gulati et al., 2019) potential Trust bias in technology across gender, generations, and culturally diverse countries like Mozambique, Afghanistan, Malaysia, Portugal, Brazil, Estonia, Cape Verde. With the survey results, we seek to establish a methodological framework beyond current disciplinary perspectives on how to design trust

countermeasures to support human-autonomy computer interactions (HCI). The second aim (2) is to demonstrate the impact of those countermeasures on supporting the trust dynamic. If successful, the results will reveal potential sources of trust influence in social media among underexplored populations. As well as will provide insights on how to deploy trust countermeasure (e.g., socially-designed, content and visual cues, visual and physical features such as voice and personality) to support Trust in human-computer interactions (HCI). This research builds from the rationale that Trust is present in nearly every human relationship, even when related to an autonomous technological artifact. The main assumptions behind this project acknowledge that current and future technology are inevitably transforming our social interactions and implicit values across culture. Thus technologists need additional insights on the influence of the Trust dynamic in fostering positive human-autonomy computer interactions. The research questions that guide our investigation are:

- RQ1: Can the Human-Computer Trust scale (HCTS) be used to define trust behaviors across different cultures?
- RQ2: How can the set of trust behaviors (across different cultures) be useful to support human-computer interactions (HCI)?
- RQ3: How to leverage user's Trust in a system?

[T] Dr. Katia Sycara, Carnegie Mellon University, katia@cs.cmu.edu
Co-Investigators: Michael Lewis, U. of Pittsburgh; Nilanjan Chakraborty, Stony Brook

Trustworthy Human Interaction with Robotic Swarms

Who Do You Trust?: Degradation, Repair, Trust and Reliance in Human Supervision of Swarms

Human understanding and inferences about the performance of swarms is limited and contaminated by factors such as compactness which itself interacts with heading variance to produce judgments of performance which differ from objective measures. In the past year we have conducted three studies investigating human ability to discern, adjust trust, and make supervisory decisions in monitoring swarms with failing robots. The first extends work on the effects of a swarm healing algorithm on supervisor trust reported last year for a single trajectory toward a goal. In the current study robots traverse way points. The effects of both failures and self-healing are made more difficult to detect because the system is continuously in transient states. Despite these difficulties results are consistent with the earlier experiment with supervisors expressing more trust for swarms without failures and an intermediate level of trust for swarms with repairs. A second experiment looked more closely at the percentage of failing robots needed to reduce trust, finding trust and intention to rely declining when 20% or more of the swarm exhibited degraded performance. The final study extended this work by giving participants the ability to intervene to change the swarm's heading to compensate for perceived failures. While participants' ratings of trust and intention to rely were similar to those in experiment 2 the decision to actually rely/intervene was unrelated to either trust or intention to rely. We discuss these findings and implications for human supervision of swarms.

[T] Lt Col Chad C. Tossell and Dr. Ewart J. de Visser, Warfighter Effectiveness Research Center, US Air Force Academy

Human & Intelligent-Agent Trust for Ethical and Effective Teaming

Given the world-wide race to incorporate AI and related intelligent technologies throughout society, militaries are seeking to harness the potential of these technologies to enable strategic advantage. Our previous work has aligned our research with two distinct research thrusts to help integrate intelligent technologies effectively as they become more ubiquitous in military contexts. With the AI-as-tool paradigm, we plan to study high performing human-autonomy teams in risky (e.g., Tesla Model X) and safe environments (e.g., flight simulators). With the AI-as-teammate paradigm, we will develop a series of studies to assess the influence of socially-intelligent and ethical mission assistants, advisors, and authorities (such as robots and virtual agents) in future conflict environments developed as part of this grant. Trust development, maintenance, and repair as well as the influence of the technology on future users will be assessed in real-world environments. The Warfighter Effectiveness Research Center (WERC) at the United States Air Force Academy is in a unique position to conduct this type of research. Our research is facilitated through the cadet capstone process as well as supported by dedicated research personnel, a well-equipped laboratory, and world-class collaborating faculty around the globe. This work will have three primary and unique benefits. The first is a strong scientific contribution to the growing field of human-machine teaming research leveraging our existing capabilities and external validity provided by our participants (i.e., cadets). The second is a valuable educational experience by involving cadets early in the research process and preparing them with AI tools and teammates to be effective future leaders of the Air Force. Last is the creation of a scientific and collaborative hub to connect academic researchers together at the academy through collaborative research efforts as well as hosting of review meetings. We plan to combine the educational experience of cadets and the collaborative nature of this research by establishing collaborations and student exchanges between research groups at Universities, Research Institutions, and Military Academies.

[T, Y] Tom Williams, Colorado School of Mines

Calibrated Norm Violation Response in Human-Machine Teaming

Robots have been demonstrated to wield significant persuasive power, and to wield moral influence over human teammates. This gives robot designers a unique responsibility to ensure that their technologies wield this influence in a positive manner. Specifically, robots, especially language-capable robots, need to avoid condoning inappropriate commands and assertions, and need to actively push back on inappropriate commands and assertions. Moreover, robots need to do so in a way that is tactful -- reinforcing important moral norms without being perceived as so aggressive and rude that they trigger reactance.

In this talk, I will describe the first year of work on a research project aiming to enable robots to automatically generate natural language responses that appropriately push back on inappropriate commands and assertions. Specifically, I will talk about (1) a new theory of robotic social agency that undergirds our research philosophy; (2) novel computational work that enables robots to avoid morally

problematic clarification requests; and (3) preliminary work arguing how robot social identity performance may be leveraged to challenge existing but problematic norms.

[T] Dr. X. Jessie Yang, University of Michigan

Trust Building in Human-Autonomy Teaming: A Reinforcement Learning Approach

Model-based and model-free approaches for modeling trust and dependence in human-autonomy teaming

As autonomous and robotic systems become more capable in perception, planning, learning, and action, there is an increasing possibility that they will become full-edged team members. The humans and autonomous agents are expected to work as a team in environments subject to uncertainty and dynamic changes. To enable effective teaming, trust has been identified as one central factor. Our project aims to develop algorithms that enable the autonomous agent to infer the human's objectives and moment-to-moment trust, and to adopt different interaction strategies for building trust and enhancing team performance.

In this talk, we will present our main results on trust inference, including a model-based approach and a model-free approach. Our proposed model adheres to three properties of trust dynamics characterizing human agents' trust development process de facto and thus guarantees high model explicability and generalizability. We will present preliminary evaluation results of the model using data collected from an ongoing human-subject experiment. In the experiment, a team comprising of a human agent and a robotic agent performs a high-workload time-critical intelligence, surveillance, and reconnaissance (ISR) mission. The robot can sense the state of the world through its sensors and perform decision-theoretic planning. Furthermore, we will describe a novel model-free approach for trust inference using Gaussian process and Bayesian optimization, and we will also present simulation results.