Trust Research in AFRL

Joseph Lyons, PhD

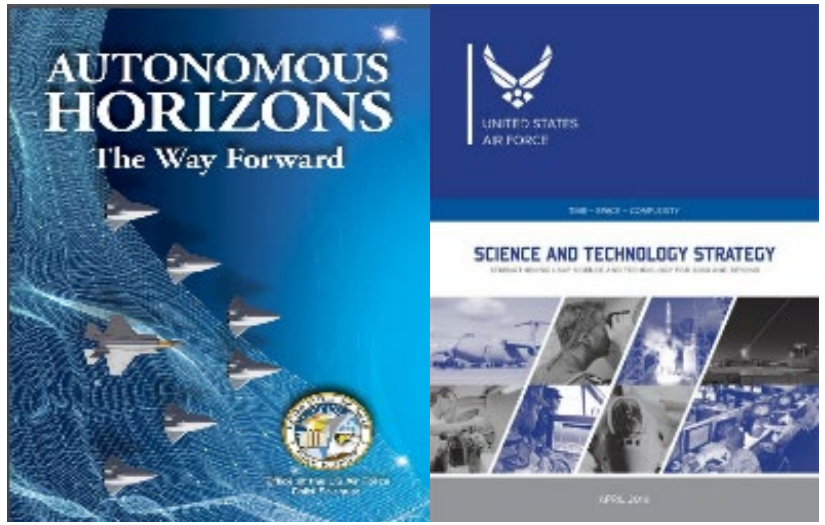
Collaborative Interfaces and Teaming Branch

29 JUL 2021

Trust Research is a Strategic Imperative

AF2030 S&T Strategy - Trusted data, Trusted AI, Trust required to support lethal combat operations

Autonomous Horizons Vol. 2 – "Autonomous systems should…Ensure trust…tenets of trust include…transparency for decision making"



Interfaces for Applied Systems
-Medusa C2 – applying Play Calling approach in novel displays
-Skyborg – Transfer of Authority of Groups/Fighter-based control

Trust/Transparency in AI – DARPA ACE; Squad-X; Alias; Trust of ML, F-35 AGCAS

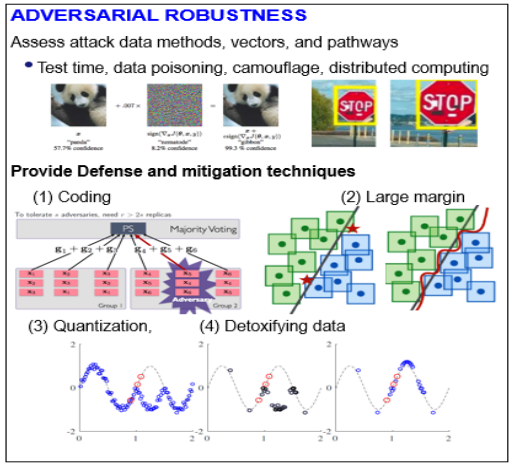
"Trust in distributed teams", "Multi-domain collaboration" – **SAB Study on Technologies for Enabling Resilient C2 (2018)**

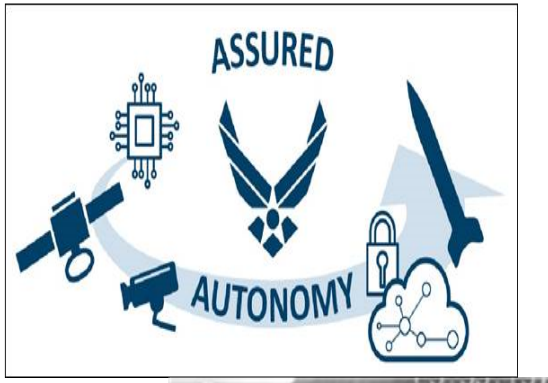
-JADC2 Operating Concept – Decision Making/Convergence of effects

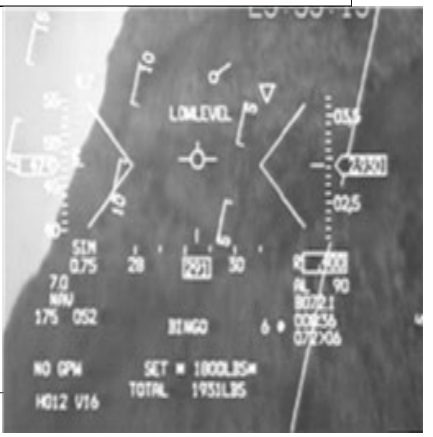
-Space Trusted Autonomy – distributed comm, mixed initiative work, trustworthiness

Trust is Relevant Across AFRL

- 711 HPW
  - Trust in autonomy, transparency, biases
- AFOSR
  - Trust and Influence Portfolio, Formal Verification Methods
- Information Directorate (*recent Trusted AI event)
  - Robust and resilient machine learning
- Aerospace Systems Directorate
  - Certification of autonomous systems/vehicle behaviors
- Space Systems Directorate
  - Space trusted autonomy
- Munitions Directorate
  - V&V to build trust in Networked Collaborative Autonomous Weapons
- Materials Directorate
  - Trust in robotics/precision manufacturing
- Sensors Directorate
  - Trusted data, data fusion





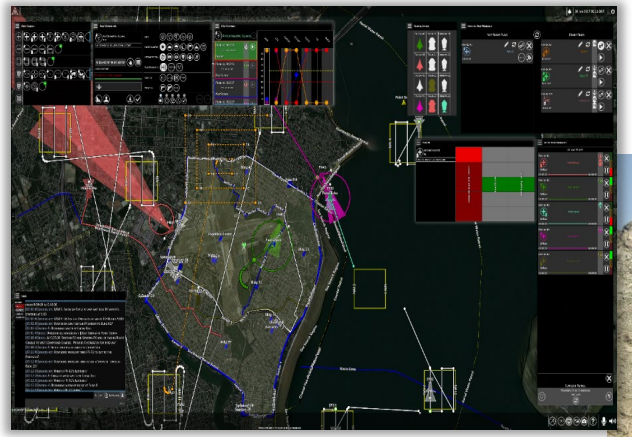


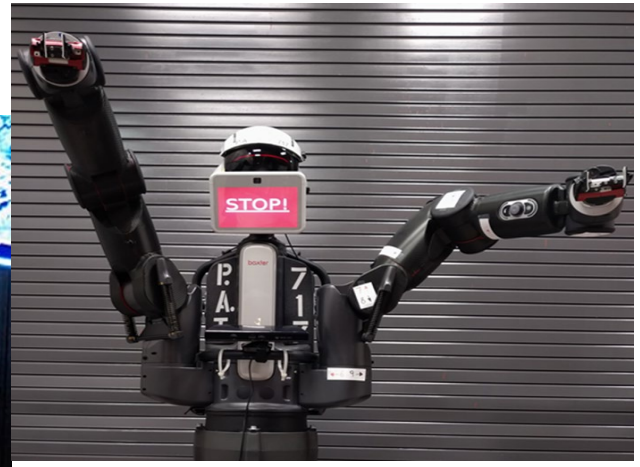
Trust Workshops/Organizing Activities (RH & AFOSR)

- Trust management in cyberspace – 2009
- Trust-based vulnerabilities 2009
  - Individual differences (PAS; suspicion)
  - Culture
- Directorate Trust Deep Dive – 2010
  - Measurement, predictors of trust, culture
- Trust measurement – 2010
- Trust in Autonomous Systems – 2012
  - 40 experts across academia-industry-gov't
  - Basic Research Initiative & multiple grants
- Trust in human-agent teams – 2015
- Support to other programs: IARPA, DARPA, OSD, NASA, FAA, IDA, ACC, etc.

Shallow Dive into 711 HPW Trust Research

- Trust in automation
  - Transparency, reliability
- Trust of fielded systems
  - Acceptance of automation
- Trust in human-autonomy teams/Robots
  - Transparency, teaming factors
- Trust of software code
  - Predictors of re-use
- Interpersonal trust
  - Swift Trust in JADC2 teams



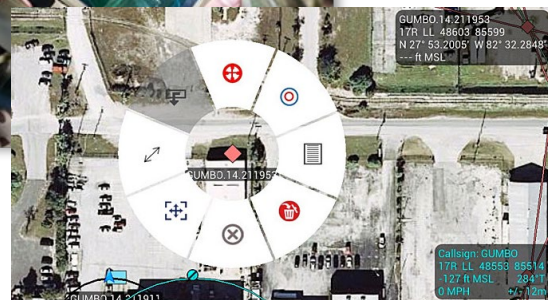


Collaborative Interfaces and Teaming CRA

Human-Autonomy Collaboration

Distributed, Heterogeneous Teaming Solutions

The Collaborative Interfaces and Teaming CRA focuses on: 1) flexible, directable, and transparent Human-Autonomy Teaming (HAT) solutions, 2) the science of human-human teaming in distributed multi-domain contexts, and 3) development of technologies to facilitate shared authority of autonomy and common ground within and between mixed human-autonomy teams.



Increased focus on teaming is an intentional strategic pivot toward JADC2

Human-Autonomy Collaboration LOE

- 6.2 FY21 Tasks
  - Models/Metrics for Human-Autonomy Teaming — Lyons et al. (2021) Frontiers in Psychology
  - Collaborative Interfaces Research
    - HMI design
    - Context-aware agents
    - Task manager

    SkyFlagONE [ABMS]; Medusa C2 [PEO Digital]; Assured Base Operations
  - Trust in Intelligent Machines
    - Swarms
    - F-35 AGCAS
    - Trust of Robots/Agents

    CMU Center of Excellence

    F-35 JPO; OSD Safety Office
  - Manned-Unmanned Teaming (Fighter-based control) — Aid for Rapid SA Acquisition
  - Transfer of Authority – distributed OPs — Battle Card Concept /AISC
  - Transparency in Machine Learning Systems (*New area in FY21)
  - Synthetic Teammates and their impact on trust in multi-team systems (*New area in FY21)

Distributed, Heterogeneous Teaming Solutions LOE

- 6.2 Tasks
  - Team performance metrics
  - Team kickstarter methods
    - How to facilitate swift trust among team members
    - Skill/role deficiencies (*New area FY21)
  - Multi-domain teaming
    - Play calling approaches for cyber
    - Multi-domain Course of Action (COA) generation and analysis
    - Integration of effects for Air, Space, and Cyber (aspirational)
  - Team resilience methods
    - How to ensure effective team process amid degradation/change (FY22 start)

Tolston et al. (2019). *BRM* (Wing Top 10 Publications 2020)
-Analysis of Black Skies Exercise Data

Capiola et al. (2020) JCEDM

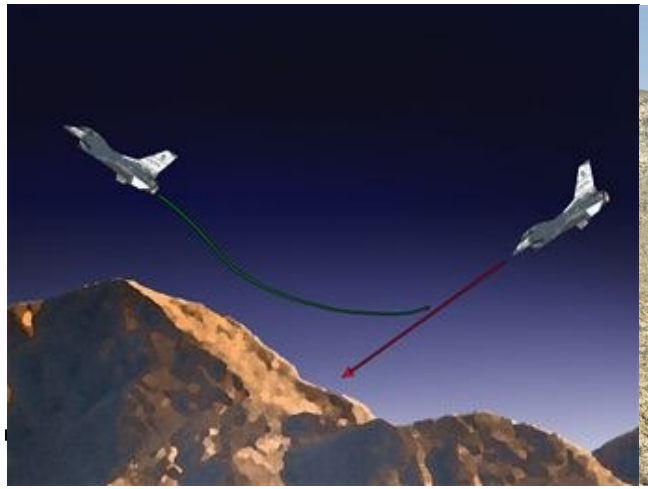
Tech Sprint 2021

805th Combat Training Squadron – Shadow Ops Center Network (ShOC-N); Nat Space Defense Center

EMS&A – integration of Cyber into Battle Management

Trust Research Example – AGCAS Study 2015-2018

- Objective: Understanding antecedents of automation trust among pilots over time
  (Ho et al., 2017a; b; Koltai et al., 2014; Lyons et al., 2016; 2017)
- Interview & survey research
  - Sampled approx. 500 F-16 pilots, 70 F-22 pilots *by ACC request (only F-16 data reported herein)
  - Baselined trust, identified trust antecedents
- Results were used to improve the system by working with engineers
- Activation data provided to various stakeholders
  - Understanding of activations outside of test were largely unknown
- Similar Studies conducted with Rotary Wing Aircraft community & F-35



AGCAS Year 1

- Year 1: Survey (N = 142), Interview (N = 168)
- Primary Findings:
  - Trust was highly variable – moderate at best
    - Uniquely related to PAS, perceived benefits, & performance
    - Pedigree of the test community was high & that helped
    - Business case was very strong
    - Chevrons were viewed very positively – helped make the system predictable
    - Activation base rates were low approx. 10%
  - Lots of early errors causing uncertainty
    - Key was attribution & technical information
  - Lots of stories – good and bad
    - Early system save was strong trust booster
  - Pilots had little knowledge of AGCAS – lots of confusion
  - Policies/practices were varied
    - Some units flew with it off, turned off for BDC/formation flights

AGCAS Year 2

- Year 2: Survey (N = 100), Interview (N = 131)
- Trust was moderate to high
  - Performance was the key driver
  - Stories of the saves pervaded the pilot community
    - Tipping point was student save w/video
  - Knowledge increased – system began to become predictable
    - Experiences with activations increased
  - Chevrons became predictable
  - Business case unquestionable
  - However, growing concern over novel nuisance factor
  - Activate rates were around 20%
  - Pilots were instructed to use PARS in training
    - Added familiarity

AGCAS Year 3

- Year 3: Survey (N = 77), Interview (N = 103)
- Trust was very high and universal
  - Saves were very well known
    - Student save video became kind of legendary
    - Student and instructor save sealed the deal – video impact
  - Perceived benefits were universal and huge trust booster
  - System was understandable and integrated into the pilot curriculum
  - Chevrons incorporated into Strafe training & ops
  - Activation rates were approx. 34%
    - Direct experience with the system was growing
    - Plus use of PARS supported experience of maneuver
      - Also use of PARS operationally that boosted AGCAS trust
  - Nuisance issue had a fix coming
    - It was understandable, predictable, and pilots had directability
  - Instructor pilot anecdote

Recent Basic Trust Research – funded under AFOSR's Trust & Influence Portfolio

- Trust biases in HRI (PI: Dr. Gene Alarcon)
  - Studies empirically examining trustworthiness biases toward robots
    - Benevolence/integrity violations (published in Applied Ergonomics)
    - Full ABI manipulations robot vs human (IEEE HMS conference; multiple manuscripts under review)
    - Effects of Perfect Automation Schema on biases (in progress)
- Human-agent teaming/Compliance (PI: Dr. Gregory Funke)
  - Capacity to cooperate in human-agent interactions (online data collection complete – new start)
  - Robot compliance (Frontiers in Psychology 2021; HFES 2019)
- Transparency in HRI (PI: Dr. Joseph Lyons)
  - Studies examining facets of transparency in autonomous robot contexts
    - Stated Social Intent (completed, published in Human factors 2021; Applied Ergonomics 2020)
    - Decision authority (completed, under review)
    - Robot etiquette (completed, under review)
- Mental models in HRI (PI: Dr. April Rose Panganiban)
  - Studies examining how mental models develop for robotic partners
    - Examined the impact of supportive communications in Loyal Wingman Scenario (published in JCEDM 2019)
    - Individual differences in trust in autonomous partners: Implications for Transparency (IEEE Transactions on Human-Machine Systems, 2020);
    - Trust in the Danger Zone: Individual Differences in Confidence in Robot Threat Assessments (submitted to Frontiers in Psychology)

POC

Any questions?

Joseph Lyons, PhD
Principal Research Psychologist,
Collaborative Interfaces and Teaming

Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Collaborative Interfaces and Teaming Branch
(711HPW/RHWC)
Wright-Patterson AFB

Phone: (937) 713-7015
E-mail: joseph.lyons.6@us.af.mil