Taylor Foster

Project ENTERPRISE Annotated Bibliography

Cañas, J. J. (2022). AI and ethics when human beings collaborate with AI agents. Frontiers in Psychology, 13, 836650.

In this paper, Cañas (2022) explores the ethical dimensions of human-AI collaboration, emphasizing the necessity of designing interactions between humans and AI agents as collaborative processes rather than treating AI merely as tools. The author argues that such collaborations require co-supervision, where both human and AI agents monitor and influence each other's actions to ensure shared responsibility for outcomes. This approach shifts the ethical focus from traditional "design goals" to the implications of joint actions and responsibilities during human-AI interactions. Cañas stresses that designing ethical human-AI collaboration requires accounting for how agents influence each other's actions. The paper identifies both cognitive and non-cognitive components of collaboration, arguing that ethical considerations should focus on the collective outcomes of joint actions rather than the isolated ethics of individual agents. This approach aligns with the concept of "collective responsibility," emphasizing shared accountability for ethically correct or incorrect results. To foster ethically sound collaboration, Cañas examines control strategies in human-AI interaction. Divergent strategies between humans and AI can lead to accidents, underscoring the need for alignment. The paper proposes two solutions: training agents to understand each other's behavior or designing AI systems to mimic human decision-making processes. Additionally, Cañas emphasizes the importance of perceptual invariants (consistent patterns used by humans in control tasks) and suggests integrating these into AI design.

Gervits, F., Fong, T. W., & Scheutz, M. (2018). Shared mental models to support distributed human-robot teaming in space. In *2018 aiaa space and astronautics forum and exposition* (p. 5340).

In an exploratory study, Gervits, Fong and Scheutz, apply the Shared Mental Model (SMM) framework in a deterministic virtual environment to demonstrate how it could benefit distributed human robot team efficiency and task performance for space operations. Gervits is an Army researcher, Fong is the Chief Roboticist at NASA Ames, and Scheutz is a Computer Science professor at Tufts University. SMMs are a mulit-domain, team management concept that relies on team knowledge, hierarchy, planning, functional roles and task environment for coordination. In their framework, equipment, task, team interaction and teammate states are stored in data structures, synchronized across agents, and implemented using representational primitives. For the study scenario, a ground-based human operator must collaborate with two robotic assistants onboard a spacecraft in lunar orbit to repair damaged interior panels. In the experimental conditions, the SMM framework was compared to a control where the robots lacked information-sharing and adaptive team behaviors. Ultimately the SMM condition had improved efficiency and performance. Although no significant reduction in workload was observed, nor was there increased operator situation awareness, they propose the SMM framework as a strategy to enable robotic interdependence and overcome the operation delays. However, formalizations in this paper cannot be validated until the SMM framework is integrated with a robotic architecture and tested in a higher fidelity environment.

Hou, Y. T. Y., & Jung, M. (2018). Robots in power. In *Proceedings of Longitudinal Human-Robot Teaming Workshop at HRI18* (pp. 325-38).

Two Cornell researchers, Hou and Jung, examine the implications of robots in powerful positions to better understand group dynamics in mixed human-AI teams. Power is understood to be the capacity of an agent to influence another party to perform an action. To this end, they argue that algorithm-based management is already influential and that intelligent non-human agents may also achieve this level of influence. Because intelligent systems can process huge amounts of data in a short period of time, Hou and Jung claim that robots would excel in leadership positions for emergency responsiveness, resolving interpersonal conflicts and creating a sense of fairness, where human biases are known to impede decision-making. They then discuss the different categories of power that originate from organization, social interaction or expertise, the latter of which is the most relevant source of power for robots to dominate. In human-robot relationships, robots that are subservient are generally less intimidating, robots that are equals are considered teammates, but the idea of robots as leaders is not well understood. Unfortunately, Hou and Jung did not share any thoughts regarding AI alignment with human values or algorithmic bias. But overall, this extended abstract lays out major questions for future robotic designers to consider in terms of hierarchy and power configurations.

Luebbers, M. B., Chang, C. T., Tabrez, A., Dixon, J., & Hayes, B. (2021). Emerging autonomy solutions for human and robotic deep space exploration*. Proceedings of Space CHI: Human-Computer Interaction for Space Exploration* (SpaceCHI 2021). Yokohama, Japan.

In a paper by researchers from the University of Colorado Boulder, Luebbers et al., autonomy challenges for deep space missions are discussed in addition to emerging technologies that could overcome the largest risks associated with using probabilistic systems. They then proceed to describe two human robot interaction (HRI) case studies that relate to spatiotemporal issues. One being an astronaut team with robotic assets on Mars and a time delay to mission control, and the other being infrequent remote contact with robots on a deep space mission to Europa. Envisioned to be complementary agents to humans on a Mars mission, robotic assets will need a shared mental model of teammate proximity, dynamic task allocation, the ability to recognize intent, self-optimization and self-evaluation for effective HRI. Meanwhile a robot deployed on Europa would need extremely flexible and explainable operation without human intervention for extended periods, autonomous goal selection, power and time allocation and fault prevention/recovery between communication events. To address these scenarios, Luebbers et al suggest future space autonomy will depend on human interpretability of complex machine learning models (Explainable AI/xAI), Virtual Augmented and Mixed reality and designing for adaptability. xAI methods include documentation datasheets, counterfactual explanation, and semantic justification. Whereas virtual reality could be used for first person point of view remote path planning or augmented reality could improve situational awareness during proximal human-AI team exploration. Lastly adaptive/adaptable control would require systems to adjust their own parameters or have human operators make the adjustments to the level of automation the robot is operating under. Ultimately this paper identifies several autonomy challenges and points to areas of active research necessary for interplanetary HRI, but with little elaboration on the specifics.

Myers, C. W., Cooke, N. J., Gorman, J. C., & McNeese, N. J. (2024). Introduction to the Emerging Cognitive Science of Distributed Human - Autonomy Teams. Topics in Cognitive Science. https://doi.org/10.1111/tops.12744

This introduction by Myers et al examines the emerging field of human-autonomy teams (HATs) through a cognitive science lens, presenting team definition, trust dynamics, autonomous teammate capabilities, and measurement metrics as four critical facets of the practice The authors establish a theoretical framework that bridges individual cognitive processes with team-level phenomena, using Newell's timescale to demonstrate how cognitive activities span from millisecond-level biological processes to month-long social interactions. The paper makes significant contributions to understanding technical integration challenges by examining how autonomous systems can effectively function as team members rather than tools, emphasizing the importance of heterogeneous capabilities and complementary skills. For safety and ethics, the paper highlights the importance of trust calibration and repair mechanisms in HATs, acknowledging that lower trust leads to decreased team performance. The authors emphasize the need for real-time measurement of trust dynamics to ensure safe and effective human-autonomy collaboration. The paper's primary limitation is its theoretical focus, with limited practical implementation guidance. However, its comprehensive framework for understanding HAT dynamics with interactive team cognition makes it an essential resource for researchers and practitioners working on autonomous system integration for collaborative activities, such as human-robot space exploration.

Nesnas, I. A., Fesq, L. M., & Volpe, R. A. (2021). Autonomy for space robots: Past, present, and future. *Current Robotics Reports*, 2(3), 251-263.

JPL technologists, Nesnas, Fesq and Volpe provide a review of space autonomy advances, assess the state of the field and justify the need for autonomy for future exploration. They enumerate the autonomous functions across mission phases before defining autonomy (only as it relates to a remote system operating independently of ground control) delineating function level autonomy (for domain-specific subsystems) from system level autonomy (whole robot operation) and establishing an autonomy continuum wherein highly prescriptive tasks and routines transitions to greater situational awareness with increasing autonomy. In their view the evolution of deep-space exploration demands more flexibility when a robotic spacecraft encounters environmental changes, and the required response time is shorter than the next communication cycle. Currently, sequence driven operations need human operators to establish situational awareness and past examples of autonomy have been constrained to in-space robotic operations, small-body proximity operations, landed missions and surface missions. Anticipating future needs for space autonomy, Nesnas et al. expect to see richer sensing and perception systems, growth in the diversity of forms robotic exploration takes, and autonomy benefiting system robustness, cost effectiveness, and productivity of in situ operations. The paper concludes by proposing a low-cost, high level autonomous SmallSat mission to a near Earth object without a priori knowledge of its exact characteristics, to test computer vision for guidance, navigation and control and surface mobility in order to further advance autonomy.

Roth, E. M., Sushereba, C., Militello, L. G., Diiulio, J., & Ernst, K. (2019). Function Allocation Considerations in the Era of Human Autonomy Teaming. *Journal of Cognitive Engineering and Decision Making*, 13(4), 199-220. https://doi.org/10.1177/1555343419878038

This comprehensive review examines function allocation methodologies in human-autonomy teaming (HAT), offering critical insights into the technical and non-technical challenges of integrating autonomous systems with human operators. The authors analyze various approaches including Cognitive Task Analysis (CTA) and Cognitive Work Analysis (CWA), demonstrating how these frameworks can address safety concerns through systematic evaluation of operational demands. The paper notably contributes to ethical considerations by emphasizing the importance of balanced function allocation that respects both human capabilities and autonomous system limitations. Regarding training implications, the authors highlight how work organization possibilities analysis can support adaptive skill development for fluid human-machine collaboration. The research particularly excels in addressing autonomy and control issues through its detailed examination of interaction requirements and the co-active design approach, which views human-machine systems as interdependent partners rather than isolated entities. The paper's thorough exploration of the function allocation trade-space provides valuable guidance for system designers in making ethical and safety-conscious decisions about automation levels. However, while the theoretical framework is strong, the paper could benefit from more concrete examples of successful implementations. This work is particularly valuable for researchers and practitioners working on the integration of autonomous systems in safety-critical environments where ethical considerations and human factors must be carefully balanced.

Tokadll, G., Dorneich, M. C., & Gonzalez-Torres, T. (2018, September). Preliminary Guidelines for Human-Agent Teams in Space Operations Beyond Low-Earth Orbit. In *2018 IEEE/AIAA 37th Digital Avionics Systems Conference* (DASC) (pp. 1-9). IEEE.

Iowa State researchers, led by Tokadl et al., explore the role of cognitive assistants (CAs) in supporting human crews operating beyond Low Earth Orbit. Their guidelines, presented through a Decision Action Diagram (DAD), aim to mitigate the impact of delayed mission control guidance and empower astronauts during off-nominal situations. The research emphasizes a human-centered approach, treating CAs as teammates activated when communication delays occur. Key points include shared decision-making functions between humans and CAs, dynamic allocation triggers, and specific levels of automation (LOA). The Decision-Action Diagram (DAD) provides a structured model for designing adaptive CAs to support autonomous decision-making. Technical integration focuses on developing adaptive autonomy systems and LOA frameworks to ensure CAs act as effective team members without excluding mission control. As for non-technical integration they recommend safety measures through fail-safe mechanisms, and robust training protocols to prepare astronauts for collaboration with autonomous systems Overall, the guidelines are preliminary and should be refined through human-in-the-loop evaluations. This research is crucial for enhancing human-agent collaboration in deep-space missions, balancing autonomy and control, and empowering space crews to handle complex scenarios independently of real-time Earth-based support.