Human Factors in Robots, Drones and Unmanned Systems, Vol. 138, 2024, 115–120
https://doi.org/10.54941/ahfe1005014

Bidirectional Human-AI/Machine Collaborative and Autonomous Teams: Risk, Trust and Safety

William Lawless

Paine College, Augusta, GA 30901, USA

ABSTRACT

We address the bidirectional challenges in developing and managing interdependence for AI/machine collaboration in autonomous human-machine teams. Recent advances surrounding Large Language Models have increased apprehension in the public and among users about the next generation of AI for collaboration and human-machine teams. The anxieties that have grown regard the risk, trust, and safety from the potential uses of AI/machines in open environments, including unknown issues that might also arise. These concerns represent major hurdles to the development of verified and validated engineered systems involving bi-directionality across the human-machine frontier. Bi-directionality is a state of interdependence. It requires understanding the design and operational consequences that machine agents may have on humans, and, interdependently, the design and operational effects that humans may have on machine agents. Current discussions on human-AI interactions focus on the impact of AI on human stakeholders; potential ways of involving humans in computational interventions (e.g., human factors; data annotation; approval for drone actions); but these discussions overlook the interdependent need for a machine to intervene for dysfunctional humans (e.g., in 2015, the copilot aboard a Germanwings airliner committed suicide, killing all aboard; in 2023, a pilot ejected from an F-35, allowing the plane to fly unguided for an additional 60 miles). Technology is advancing rapidly: Self-driving cars; drones able to fly and land autonomously; self-landing reusable rockets; Air Force loyal wingmen. The technology is available today for bi-directional AI/machine collaboration and autonomous human-machine teams to better protect human life now and in the future. Thus, despite the engineering challenges faced, we believe that the technical challenges associated with humans and AI/machines cannot be adequately addressed if the social concerns related to risk, trust and safety caused by bi-directional forces are not also taken into consideration.

Keywords: Explainability, Risk, Trust and safety, Joint awareness, Shared mental models, Systems design, Engineering and operations, Assurance, Test and evaluation, Operational system failures (e.g., loss of communication)

INTRODUCTION

In our research, which we review in this chapter (along with links to our findings and generalizations), we have developed a model of interdependence to find that for teams, the least structural entropy production, ΔSEP, and the

maximum entropy production, ΔMEP, form a tradeoff that humans are able to exploit:

$$\Delta SEP * \Delta MEP \geq C. \quad (1)$$

With this Equation (1) as a model, we have reached the following generalizations and conclusions.

1. As one of the tradeoffs reflected by Equation 1, the best teams minimize structural entropy production, SEP, to be able to maximize their team's overall performance as determined by its maximum entropy production, MEP. First, the best teams and organizations reduce redundancy in their organizations (characterized as an excess in the number of employees in a team or organization compared by their competitors doing the same work), which we found to be true for the largest oil producers in the world (Lawless, 2017a) and then which we replicated for the largest militaries in the world (Lawless, 2017b). We found in both studies that more redundancy was prevalent for teams and organizations in less free than in more free countries, suggesting that redundancy was a hidden source of corruption.
2. Based on Equation 1, we generalized tradeoffs to trust (Lawless & Sofge, 2017), to social noise (Lawless et al., 2018a), to models using Artificial Intelligence (AI; in Folds & Lawless, 2018), and eventually to deception (Lawless et al., 2018b), resulting in an editorial in AI Magazine in 2019 (Lawless et al., 2019). Before we completed this stage of our research, in an edited book (Lawless et al., 2017), we further wrote in another editorial about trust, concluding that trust was bidirectional, a startling result that has played a substantial role in our research (briefly discussed with two examples in the abstract and with an example near the very end of this chapter).
3. That finding led to an invited article on autonomy that became a bet based on a hunch, predicated partly on the successful exploitation of bidirectional trust, that human-machine teams were close to achieving autonomous operations that might occur sometime during the next five years (Sofge et al., 2019).
4. Our next generalization was to deception (Lawless, 2019; Lawless et al., 2020). For deception to operate or do work, we reflected on Equation 1 to theorize that for a deceiver in an organization to be successful, it (human or machine) must be able to minimize the entropy generated by its presence compared to others in the team or organization by not allowing its presence to increase a team's structural entropy production (SEP) until its deception had served its purpose.
5. Our subsequent and third generalization was to recognize the similarity, based on state dependency, between interdependence for human-human interactions and entanglement at the quantum level (Lawless, 2020). Equation 1 is based on this premise; we have also included emotion as a model of the state of a team when its state is elevated above the team's ground state. Hindered by its validation crisis, one of the problems associated with the social sciences is the inability of social scientists

to successfully generalize their research findings (e.g., the invalidity of self-esteem; implicit racism; ego depletion), leading to a complete failure with generalization (e.g., honesty; superforecasting; etc.). We further concluded that studies of teams, including even with live video, cannot fully capture the interaction in order to be able to recreate the interaction. In theory, this inability occurs principally because social science is predicated on the use of independent and identically distributed data (i.i.d. data, namely Shannon information) which, by definition, cannot recreate states of interdependence.

6. Next, working with Nancy Cooke (Cooke & Lawless, 2021), we began to suspect that something was more important than intelligence for the best performing teams. She had concluded that intelligence was in the team's interactions, not a precondition for the best performance of a team's interactions. We extended Cooke's finding to team structure, a breakthrough, and discussed next.
7. Mindful of Cooke's work, we began to review prior research only to reexamine the importance of structure to decision-making over time (Akiyoshi et al., 2021). We reviewed the problems associated with consensus-seeking compared to majority rule decisions; we concluded that consensus seeking increases redundancy by giving too much power to a minority in a team or organization to be able to block a decision (which allowed us to rename consensus seeking as minority-rule decision making in contrast with majority rules), thereby not allowing a team to process the information available to it. During decision-making in a competitive environment, the winning team must be able to defend itself while it seeks vulnerabilities in its opponent. Consensus seeking in a team generates too much internal entropy (e.g., with conflict or disagreement), making consensus-seeking rules unable to be productive by, in our case study, accelerating the cleanup from the mismanagement of radioactive wastes (Akiyoshi et al., 2021). Second, replacing members of the best teams is indicated by their fittedness; i.e., a good replacement that fits the existing team structure is indicated by a reduction in structural entropy (Lawless et al., 2023b).
8. Vulnerabilities were also discovered in one's own team and in an opponent's team by an increase in structural entropy production, SEP; a decrease in performance of a team's maximum entropy production, MEP; or both (Lawless, 2022c). To establish this finding, we analyzed data generated and provided by the United Nations for all of the Middle Eastern North African (MENA) nations. We found that the better and more widely educated was the population of a country, the freer were its people, the better became its ability to innovate, and the less corruption that was experienced overall by the country. Of the MENA nations, Israel was the leading country across all categories.
9. In our most recent research, we focused on the relationship between time, energy and entropy. First, we generalized Equation 1 to uncertainty in time and energy. Then for the GDP of the top countries listed in the United Nations, we compared time in starting a new business, energy consumed per capita, freedom, innovation and other factors (Lawless, 2024; under review but a preprint is available). Surprisingly, we found that the more

Shannon information generated by a nation, the more successful it was (namely, generalizing for the Shannon information specifically generated by checks and balances; e.g., reporting on political and business conflicts with a free press). We found that time is a critical factor in the best run teams and most innovative organizaqtions, and that the least well run organizations and nations suppressed interdependence.

Human-Machine Teams: Conclusions and the Path Forward

In this section, we provide two examples of bidirectionality before we move on to review the conclusions from our research and the path forward that we are proposing.

The first example is the failure of bidirectional interdependence to be established between a human and a machine in a team. The first pedestrian fatality was by an Uber car in 2018. The Uber car was experiencing difficulty categorizing a pedestrian in front of the Uber car who was crossing the road late at night but outside of the pedestrian walkway. In that the Uber car and its human operator were acting independently of each other, and in that the operator may have been distracted, action by the operator to prevent the fatal accident was not sufficiently timely to prevent the fatality (Lawless et al., 2023b).

As the second example of applying bidirectional interdependence to the case of joint awareness between a human and its AI collaborators or teammates, even if a machine executes its role well, a scientist, engineer or ethicist might address whether a machine needs to be "aware" of what it is hearing from its human teammate, whether it is sufficient for a machine to nod on occasion during a conversation to act as if it is aware, or whether the machine must be able to do both.

CONCLUSION

The solution to how best to improve the human interaction, and by generalization to human-machine teams, is entropy. Interactions are structured for a purpose, in large measure to be repeatable. When an interaction structure wastes energy to produce more than minimum structural entropy production (i.e., when SEP is not at a minimum), less energy is available for the interaction to be productive. An excellent example is an argument over the prices published in a store; teammates constantly fighting; or a married couple who dislike each other yet are still trying to raise children. To be able to achieve maximum performance, SEP must be minimized. Once SEP has been minimized, because energy has not been wasted on structure, the team or organization has a chance to be able to direct all or most of its available energy to maximize its productivity (i.e., MEP). The possibility of achieving MEP sets the stage for higly uncertain situations, as exists diuring a competition, formal decision-making, or war, where teams must make a decision on the best path forward by debating or fighting each other. Debates are best between opposing (orthogonal) viewpoints (Lawless & Moskowitz, under review).

In our last editorial (Lawless et al., 2023a), we briefly reviewed our findings and the research path forward. In that editorial, we included the value of boundaries to link our research with Herbert Simon's bounded rationality, to Nash's countering to achieve an equilibrium, and to the court system facing uncertainty in its decision path going forward. For example, war fighters attempt to control the air space over the field of battle; transportation engineers use roundabouts to make intersections safer; and the court system operates inside of a boundary set behind closed doors. This allowed us to see a way to recover a limited rationality in decision making, but only when that rationality was embodied and not strictly cognitive (i.e., not disembodied, tacit or both).

In closing, all interactions have impediments (prior rules of engagement; government laws; business rules; social and religious rites; etc.). Minimizing these impediments, like redundancy, allows more of the energy available to let a team or organization maximize its productivity. A surprising conclusion oft repeated is that by suppressing interdependence, an authoritarian run organization (team, business, government) reduces its decision advantage in competitive situations (Lawless, 2024).

ACKNOWLEDGMENT

The author would like to acknowledge his summer research opportunities at the Naval Research Laboratory in Washington, DC, that have been funded almost every summer since 1993. Since 2012, he deeply appraciates the help of his current mentor, Ranjeev Mittu (Code 5580), Information and Decision Sciences Branch, Information Technology Division, U.S. NRL, 4555 Overlook Avenue, SW, Washington, DC 20375.

REFERENCES

Akiyoshi, M., Whitton, J., Charnley-Parry, I. & Lawless, W. F. (2021), Effective Decision Rules for Systems of Public Engagement in Radioactive Waste Disposal: Evidence from the United States, the United Kingdom, and Japan. In Lawless, W. F., Mittu, R., Sofge, D. A., Shortell, T. & McDermott, T. A., Systems Engineering and Artificial Intelligence. Springer, Chapter 24, pp. 509–533. doi: 10.1007/978-3-030-77283-3_24.

Cooke, N. J. & Lawless, W. F. (2021), Effective Human-Artificial Intelligence Teaming, In Lawless, W. F., Mittu, R., Sofge, D. A., Shortell, T. & McDermott, T. A., Engineering Science and Artificial Intelligence, Springer.

Folds, Dennis & Lawless, William (2018, 12/3), Naval Research & Development Enterprise (NRDE). Applied Artificial Intelligence (A2I) Summit. An Anthology of the Dialog of the Summit. San Diego, CA. October 15–19, 2018.

Lawless, W. F. (2017a), The entangled nature of interdependence. Bistability, irreproducibility and uncertainty, Journal of Mathematical Psychology, 78: 51–64.

Lawless, W. F. (2017b), The physics of teams: Interdependence, measurable entropy and computational emotion, Frontiers of physics. 5:30. doi: 10.3389/fphy.2017.00030.

Lawless, W. F. & Sofge, D. A. (2017), The Intersection of Robust Intelligence and Trust: Hybrid Teams, Firms and Systems. In Lawless, W. F. Mittu, R., Sofge, D., & Russell, S. (Eds.), *Autonomy and Artificial Intelligence: A threat or savior?* New York: Springer, pp. 255–270.

Lawless, W. F. Mittu, R., Sofge, D. & Russell, S. (Eds.) (2017), *Autonomy and Artificial Intelligence: A threat or savior?* New York: Springer.

Lawless, WF., Wood, J., Stachura, M. & Wood, E. A. (2018a), An application of interdependence theory to military medical research teams: Cultural noise, tradeoffs, and meaning, Journal of Enterprise Transformation, doi: 10.1080/19488289.2017.1419318; available from http://dx.doi.org/10.1080/19488289.2017.1419318.

Lawless, W. F., Mittu, R., Moskowitz, I. S., Sofge, D. A. & Russell, S. (2018b), Cyber-(in) Security, Context and Theory. Proactive Cyber-Defenses, in Lawless, W. F., Mittu, R. & Sofge, D. A. (Eds.), Computational Context: The Value, Theory and Application of Context With Artificial Intelligence. CRC Press.

Lawless, W. F., Mittu, R., Sofge, D. A. & Hiatt, L. (2019a), Editorial (Introduction to the Special Issue), "Artificial intelligence (AI), autonomy and human-machine teams: Interdependence, context and explainable AI," *AI Magazine,* 40(3): 5–13. https://doi.org/10.1609/aimag.v40i3.2866

Lawless, W. F. (2019), The Interdependence of Autonomous Human-Machine Teams: The Entropy of Teams, But Not Individuals, Advances Science, *Entropy* 2019, *21*(12), 1195; https://doi.org/10.3390/e21121195.

Lawless, W. F., Mittu, Ranjeev, Moskowitz, Ira S., Sofge, Donald & Russell, Stephen (2020), Cyber-(in) security, revisited: Proactive cyber-defenses, interdependence and autonomous human-machine teams (A-HMTs). In P. Dasgupta, J. B. Collins & R. Mittu (Editors), Adversary aware learning techniques and trends in cyber security. Switzerland: Springer Nature.

Lawless, W. F. (2020), Quantum-Like Interdependence Theory Advances Autonomous Human–Machine Teams (A-HMTs, *Entropy*, *22*(11), 1227; https://doi.org/10.3390/e22111227.

Lawless, W. F. (2022c), Interdependent Autonomous Human-Machine Systems: The Complementarity of Fitness, Vulnerability & Evolution, Entropy, 24(9):1308. doi: 10.3390/e24091308.

Lawless, W. F., Sofge, Donald A., Lofaro, Daniel, & Mittu, Ranjeev (2023a), Editorial: Interdisciplinary Approaches to the Structure and Performance of Interdependent Autonomous Human Machine Teams and Systems, Frontiers in Physics, eBook, retrieved 3/1/2023 from https://www.frontiersin.org/articles/10.3389/fphy.2023.1150796/full.

Lawless, W. F., Moskowitz, I. S.; Doctor, K. Z., A Quantum-like Model of Interdependence for Embodied Human–Machine Teams: Reviewing the Path to Autonomy Facing Complexity and Uncertainty, *Entropy*, *25*, 1323, (**2023b**). https://doi.org/10.3390/e25091323

Lawless, W. F. Time, Entropy and Shannon Information: Toward the Evolution of Autonomous Human-Machine Teams. *Preprints* **2024**, 2024011653. https://doi.org/10.20944/preprints202401.1653.v1

Lawless, W. F. & Moskowitz, I (under review), Shannon Holes, Black Holes and Knowledge: Can a Machine become a "Self-Aware" Teammate?

Sofge, Donald (Referee), Mittu, Ranjeev (Con Bet) & Lawless, W. F. (Pro Bet) (2019), AI Bookie Bet: How likely is it that an AI-based system will self-authorize taking control from a human operator? AI Magazine, 40(3): 79–84.