Balancing Depth vs Breadth Search Behavior in Explainable AI

Artificial intelligence (AI) systems have decision-relevant applications across a range of domains, from financial services [10] to health care [4] to counterterrorism [5]. In each domain, users require an understanding of the AI system's decision-making rationale to justifiably rely on the system for the support or execution of their operations. As users cannot be counted on to have a technical command of AI, research on "explainable AI" aims to augment the output of an AI system with automated generation of an explanation of the system's decision processes that is human understandable and useful to the intended audience [1, 3].

Explainable AI constitutes a core component of inward-facing frameworks for the ethical deployment of autonomous weapons systems (AWSs). The debate on AWSs is generally outward-looking, addressing how such systems might have adverse impacts on non-combatants, adversaries, or broader social, political, legal or military goals [9]. However, inward-facing concerns exist, specifically regarding the adverse impacts on warfighters that deploy with AWSs [10]. Combat requires more than any other domain that AI systems are transparent in their decision-making to ensure human-machine interoperability and ethical mission completion. Explainable AI is tasked with providing this understanding, absent of which "AWSs ought not, and actually will not, be deployed in battlefield environments" [9, p.2].

Defining how this task should be executed means facing a subtle but significant obstacle. The information demand of warfighters is likely to shift throughout combat. Therefore, the information supply of the explainable AI has to continuously adjust to re-establish the equilibrium needed for optimal understanding of the AWSs. In practice, the optimal explanation of the AWSs may be descriptive at time t, causal at time t+1 and counterfactual at time t+2. If the information provided by the explainable AI fails to meet equilibrium demand (i.e. by always providing causal explanations), the warfighters face a choice between lowering their trust in the AWSs (as they cannot fully understand it) or compromising their information demand to adjust to the supply offered by the explainable AI. Both of these choices lead to suboptimal outcomes.

As a result, the changing configuration of information demand and supply throughout combat poses a problem for enforcing rigid thresholds of explainability on AWSs deployment. Concurrently, serious moral concerns arise about weakening current rigid thresholds of causal explainability [9, 10]. Within these constraints, an understanding of AWSs is best provided by adaptively balancing two opposing approaches to explainable AI in battlefield environments.

The first approach posits a rigid threshold of causal explainability ("Why did the system do X?") [9]. The second approach posits a weaker threshold of descriptive explainability ("What did the system do?") [2, 6, 7]. Both approaches arrive at their respective threshold based on different assumptions about decision-making in battlefield environments. This difference is best understood in graphical form (Figure 1). The first approach assumes an environment that consists of a small set of decision paths that are relatively deep, with one path consisting of multiple

decision points (nodes). The second approach assumes an environment that consists of large set of decision paths that are relatively shallow, with one path consisting of few decision points.

Each environment yields a different optimal solution for maximizing informational gain. For the first approach to explainable AI, information is assumed to be vertically centered. Informational gain is maximized by going down a path first before switching across to the next path. For the second approach, information is assumed to be horizontally centered. Informational gain is maximized by going across paths first before going down a path. Adapting the names of common search algorithms, the optimal solution of the first approach can thus be classified as "depth-first search", the optimal solution of the second approach as "breadth-first search".

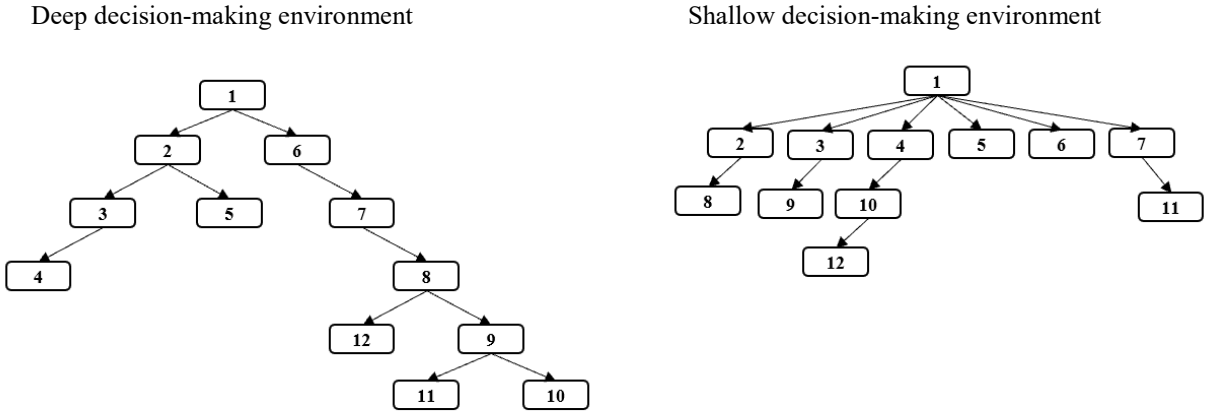

Figure 1. Example drawing of deep (left) and shallow (right) graphs to contrast the different assumptions about decision-making in battlefield environments. Numbers at each node represent the sequence of searching through the environment to maximize informational gain. In the left graph, information is vertically centered and searched best through a "depth-first search" that goes down paths before it goes across paths. In the right graph, information is horizontally centered and searched best through a "breadth-first search" that goes across paths before it goes down paths. A similar figure is given in [8] in the context of information foraging in real time strategy games.

Consequently, the thresholds of each approach to explainable AI reflect its respective optimal search behavior. Causal explainability emerges as a threshold in an environment with few possible decision paths but each with a deep composition of causally related decision points. Descriptive explainability emerges in an environment with many possible decision paths but each with a shallow composition of causally related decision points. In the first environment, warfighters can and need to focus on how a specific decision came about. In the second environment, warfighters simply need to keep track of which of multiple decisions was made.

The information environment of combat changes continually [1, 7] and may require warfighters to transition rapidly between shallow and deep decision-making. The explainable AI of AWSs needs to adapt to the corresponding changing information demand. Therefore, the explainable AI should adaptively balance depth-first and breadth-first search behavior in providing information. This entails that the thresholds for explainability of AWSs cannot be rigidly set as deep and shallow environments do not share a common threshold for optimal explainability.

Policymakers need to carefully address this matter. The ability of warfighters to understand AWSs to the extent that they can build "deep trust" [9, p. 2] into the systems is a central component of the ethical and accountable use of AWSs. However, so is the enforcement of a rigid threshold for explainability of AWSs. This leaves policymakers with two options. First, enforce a rigid threshold for explainability and accept that his may disrupt warfighters understanding of the AWSs. Second, accept that the threshold for explainability cannot be rigid but ensure that warfighters' understanding of the AWSs is maintained throughout combat. Policymakers need to be aware that the choice between both options is a choice they need to make to ensure that future combat is conducted in adherence to the highest ethical standards.

Leo Klenner

References

1. T. Besold and S. Uckelman. 2018. The What, the Why, and the How of Artificial Explanations in Automated Decision-Making. *ArXiV e-prints*, August 2018.
2. J. Dodge, S. Penney, C. Hilderbrand, A. Anderson, an d M. Burnett. 2018. How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games. 2018. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–9 (paper no. 562).
3. D. Doran, S. Schulz, and T. Besold. 2018. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017*, Vol. 2071 of CEUR Workshop Proceedings. CEUR-WS.org.
4. J. De Fauw , P. Keane, N. Tomasev, et al. 2018. Automated analysis of retinal imaging using machine learning techniques for computer vision. *F1000 Research*, 5, 1573 (2016).
5. M. Al Hasan., V. Chaoji, S. Salem, M. Zaki. 2016. Link prediction using supervised learning. In *Workshop. on link analysis, counter-terrorism and* security (2006).
6. M. Kim, S. Kim, K. Kim, and A. Dey. Evaluation of StarCraft Artificial Intelligence Competition Bots by Experienced Human Players. 2016. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1915–1921.
7. S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, and L. Simpson, M. Burnett. 2018. Toward Foraging for Understanding of StarCraft Agents: An Empirical Study. In *Proceedings of the 23d International Conference on Intelligent User Interfaces*. ACM, 225 – 237.
8. HM. Roff. 2015. Lethal Autonomous Weapons and Jus Ad Bellum Proportionality. *Case Western Reserve Journal of International Law*, 47 (Spring 2015), 37–51.
9. HM. Roff and D. Danks. 2018. "Trust but Verify": The Difficulty of Trusting Autonomous Weapons Systems. *Journal of Military Ethics*, 17, 1 (June 2018), 2–20.
10. L. Zuo, W. Long., Y. Guo. 2018. Extreme Market Prediction for Trading Signal with Deep Recurrent Neural Network. In Shi Y. et al. (eds) *Computational Science – ICCS 2018*. ICCS 2018. Lecture Notes in Computer Science, vol 10861. Springer, Cham.