30 November 2021

Meaningful Metrics for Demonstrating Ethical Supervision of Unmanned Systems

Don Brutzman and Curt Blais
Naval Postgraduate School (NPS), Monterey California USA

Abstract. Metrics for AI are important, as illustrated by the workshop topics of interest. We note that commonplace gaps in applied AI derive from "Here are the measurements we know how to take" which are too easily over-extrapolated into conclusions of interest. In other words, such precise metrics are necessary and appealing but may not broadly apply to general situations. We assert that necessary subsequent questions are "How do we define meaningful objectives and outcomes for a current unmanned system," "How do we measure those characteristics that indicate expected success/failure," and "Once we can measure meaningful results, how do we assemble exemplars into test suites that confirm successful completion across ongoing system life cycles?"

In our work, moral responsibility and authority for ethical behaviors by remote autonomous unmanned systems lies with the humans responsible for robot behavior. Lines of success or failure are clearly defined when delegating tasks to robots which have the capacity for life-saving or lethal force. Goals, constraints and metrics that are shared by humans and robots are formally verifiable as consistent and further testable in repeatable ways. This point paper explores potential design principles of broad value to ethical AI efforts.

Editorial note: this is an outline of intended discussion topics. Further details and improvements are under consideration.

1. **Metrics are essential**. Too many AI systems have ill-defined metrics that do not align with ambitious goals being pursued.
2. **Precise metrics are necessary**. All claims are suspect if they are not built upon a basis that clearly answers the design question "How do you measure that?"
3. **General metrics are elusive**. Metrics must inform the successful evaluations of objectives or else they are confusing and counterproductive.
4. **Humans own ethical responsibility and authority**. Such legitimacy cannot be fully delegated as ill-defined "autonomy" in AI systems. We are not interested in deciding whether specific activities are ethical or not, that is the responsibility of human actors carrying out shared laws and governance.
5. **Can we evaluate whether we are getting better or worse?** Once a successful system is designed, activities have meaningful measurements, and outcomes are testable, a repeatable TestDevOps test suite might confirm ongoing capabilities (and shortfalls) across different software builds, different human supervisors, and different mission goals.

We will briefly describe the following references of significant interest to workshop participants:

a. "Necessary DoD Range Capabilities to Ensure Operational Superiority of U.S. Defense Systems," National Academy of Sciences report, 2021. https://www.nap.edu/catalog/26181/necessary-dod-range-capabilities-to-ensure-operational-superiority-of-us-defense-systems
b. IEEE AI Ethics P7000-series suite of standards. https://ethicsinaction.ieee.org
c. UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence, 25 NOV 2021. https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence
d. Ethical Control of Unmanned System. https://savage.nps.edu/EthicalControl

After posing these assertions and assets to the group, we invite discussion and indeed argument over what matters most in our shared endeavors towards *Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning*.