

DOD Adopts Ethical Principles for Artificial Intelligence¹

FEB. 24, 2020

On February 24, 2020, the Department of Defense released the following five “Ethical Principles for Artificial Intelligence”. This paper suggests what they mean as far as deliverables for systems that are produced according to these principles. Measurable terms are highlighted in yellow, and footnotes are provided to suggest what they mean.

1. Responsible. DoD personnel will exercise appropriate levels of judgment and care², while remaining responsible³ for the development, deployment, and use of AI capabilities.
2. Equitable. The Department will take deliberate steps to minimize unintended bias⁴ in AI capabilities.
3. Traceable⁵. The Department’s AI capabilities will be developed and deployed such that relevant personnel⁶ possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

¹ <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/> Official Release by Department of Defense

² The term "appropriate" implies measurable. The term "judgment" suggests that the values impacting behavior must be exposed. Appropriateness can only be determined and reviewed if the behavior can be completely understood.

³ The term "responsible" suggests that the behavior of the AI needs to be traceable to groups and individuals responsible.

⁴ This highlights that systems contain bias (or measurable influencing factors). To minimize unintended bias these values must be exposed to those who are responsible.

⁵ This suggests that all influencing factors must be measurable and traceable. It must be "easy" to trace the policies that control the behavior of AI based systems or nobody will ever investigate them.

⁶ The reference to “relevant personnel”, while not specific, should suggest that that the methodology behind the AI should not be so costly to develop and maintain that it is impractical to implement.

4. **Reliable**⁷. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
5. **Governable**⁸. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

Summary:

Knowledge Enhanced Electronic Logic (KEEL) "technology" provides human-on-the-loop AI capabilities today for systems (devices or software applications) that need expert decision-making and adaptive operational control. KEEL technology (and supporting tools) supports all of the "Ethical Principles for Artificial Intelligent", in addition to making it easy to deploy AI capabilities as cognitive components that are platform and architecture independent. The cognitive components are also suitable for small real-time devices and satisfy code certification requirements.⁹

Also, one might suggest that creating explainable and auditable AI should be important to any AI-based systems for many reasons (beyond just their "ethical components"). If you cannot explain the behavior of these systems, you cannot fix them or extend them or retain control of them. If you are using AI in competitive / adversarial applications, it will be mandatory that you can change the tactics and strategies quickly. It will also be important to be able to insert new data sources into these systems as quickly as possible to maintain a competitive advantage, or to respond to new threats.

⁷ This suggests that the behavior must be deterministic and be mathematically explicit. If you cannot explain it, you cannot audit it. In addition to "well-defined uses" AI-based systems must be able to handle cases they have never been exposed to before! Otherwise it is likely that disastrous consequences will arise. It must be possible to perform after-mission reviews to ensure that the systems are performing as desired.

⁸ One might suggest that this means NO will override MUST (as an emergency stop in Industrial Automation). Again, to do this, it must be possible for humans to oversee the behavior of the AI (human-on-the-loop control).

⁹ <http://www.compsim.com/news/News%20Release%20-%20Public%20Release%20of%20NSWCDD%20Technical%20Report%20on%20KEEL.pdf> Public Release of Technical Report "Review of Knowledge Enhanced Electronic Logic (KEEL) Technology: NSWCDD/TR-16/103