Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Penn Symposium on Social Implications of Autonomous Military Systems, May 24, 2021

Toward Roboticsts' Ethics for Robots' Design and Use

D. E. Koditschek
ESE Department, University of Pennsylvania
Philadelphia, PA 19104

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Ethically Aligned Design, 1st ed. IEEE, 2019

- "Guidance for consideration" by govts, businesses, and public
- Views and opinions in collaborative work
  - authored by
    - IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS)
    - ~1000 academic, industrial, government participants
  - "do not necessarily reflect the position of their respective institutions or of IEEE" (world's largest technical professional organization)
- Document preparation
  - "open, collaborative & consensus building approach"
  - deliberative emphasis
    - initiated 2016, first edition published 2019
    - two prior versions; debated, posted & circulated, critiqued, adapted, iterated
  - effort to integrate diverse perspectives & expertise
    - polylingual, multi-cultural
      e.g. Buddhist, Confucian, Shinto, Taoist, Ubuntu, Vedic, Western perspectives
    - glossary translating terms into technical dialects
      - e.g. "agency" as defined and used in
      - computational, engineering, government/policy, philosophy literature

Context

a US Robotics Research Lab

(thanks to Prof. Lingel)

[Topping et al. IROS'17]

- kod*lab focus: the synthetic science of robotics
  - 4 Postdocs + 8 PhD students + numbers of younger students
  - theses: engineering, cs, math, biology, paleontology, …
  - collaborations: biology, geology, math, philosophy, psych
- kod*lab funding: ~ $45M from ~30 PI'd grants over ~ 40 years
  - 16 NSF grants (~13% funding); 13 DoD grants (~87% funding)
  - unclassified work; unrestricted publications; unrestricted teaching
  - restricted private international communications unrelated to funding
    - most common: DoC regulations governing "deemed exports"
    - uncommon: DoS regulations governing adversarial nations

Overview of EAD Report

- Overarching Principles
  - artifacts must
    - promote human rights & well–being
    - enhance human agency (data, identity)
    - reflect & promote *global* ethical human wisdom
  - designers' responsibilities include
    - transparency (explainable decisions) & accountability (apportioned attribution/liability/culpability)
    - evidenced effectiveness; calibrated user/operator competence; anticipated misuse guard-rails
- Results & Impacts
  - IEEE: standards committees; A/IS ethics courses; technology certification protocols
  - public exchange: global forum; AI commons;
  - public policy: connections to UN ,EU, OECD, natl. govts.

Rough Summary of EAD Contents

- Guidelines for Ethical Research and Design
  - education & research: ethics must be part of core education and practice, including
    - teaching by outside experts
    - exposure to cultural variability and norms
    - development of ethical standards
  - corporate practices: achieving values-informed profit entails both
    - internal leadership (top-down) and empowerment (bottom up)
    - as well as external oversight and certification standards
- Embedding Values in A/IS
  - normative: identification, representation, implementation of local communities' values
  - challenge: tracking variation over time and conflicts in norms, effective computational representation and implementation; graceful failure
  - evaluation: formal specification; bias identification; third-party access/verification
- Policy & Law
  - designer mandates:
    - ensure AI/S promote internationally recognized legal norms
    - focus new research to address challenges of embedding ethics/values in core technology
    - educate governments and public to create policies ensuring ethics in AI/S
  - user mandates:
    - ensure AI/S promote internationally recognized legal norms
    - acknowledge & achieve informed view of AI/S role in legal systems
    - deny legal "personhood" status for AI/S (for now)

EADv2 Chapter on LAWS

"Reframing Autonomous Weapons Systems"

- included in v2 but not in 1e "for timing reasons"
- focus on kinetic LAWS (physical harm); consideration of cyber
- emphasis on meaningful human control
  - transparent & explainable technology
    - understandable adaptive/learning components
    - predictable behaviors
  - accountable & controlled deployment
    - identifiable, responsible human operators
    - audit trails to document provenance and responsibility
  - informed designers
    - developers understand the implications of their work
    - development of professional ethical codes
    - shared concepts afford compliance with international & local law

Issues Raised in EADv2 LAWS Chapter

- Conceptual Challenges
  - Definitional Confusion
    - no clear technical understanding of “autonomy”
    - human “in/on” the loop distinction vague/inadequate relative to emerging technologies
  - Absence of professional codes of ethics
    - designers’ ethical obligations beyond legal requirements
    - professional organizations’ ability to offer practicable resources/advice to individuals
- Socio-political Challenges
  - Poor understanding/control of destabilizing/escalating risks
    - real-time: human control eroded by shrinking time constants
    - deploy-time: geopolitical arms race dynamics
    - design-time: conventions thwarted by compromised accountability/attributability
  - Ease of abuse
    - individual: easy (or intrinsic) violations of human dignity
    - local: inappropriate use by domestic police or private security forces
    - global: availability to and proliferation by non-state actors
- Technical Challenges
  - unreliability due to design complexity or scaling effects
  - unpredictability due to adaptive capacities or poorly delimited agency

Ethically Aligned Military Robotics?

- Antecedent Positions
  - US DoD Directive 3000.09 on LAWS
    - "weapons systems that once activated can select & engage targets without further intervention"
    - "allow commanders and operators to exercise appropriate levels of human judgement"
  - Proposed Bans:
    - On development: Future of Life, Human Rights Watch, Intl. Comm. Rob. Arms Control, UN Human Rights Council, …
    - On deployment but not on research & development: China
    - No ban on research, development or deployment: US, Russia
- Next Steps
  - International Engagement
    - world government negotiations seem to be stalling
    - what help might robotics offer users (govts) and their agents (ambassadors, lawyers, etc)?
    - to-do: identify concepts, terminology, emerging capabilities in need of technical definitions & standards
  - Disciplinary Maturation
    - designers' obligations expand in step with the utility of their designs
    - what help might robotics receive (history, law, philosophy, political science …) toward those obligations?
    - to-do: identify background concepts and literature needed to codify roboticists' ethical education and guideline
  - Today
    - panel discussion will hopefully begin to address such questions
    - post-symposium survey will hopefully elicit basis for working groups
  - This Year
    - disseminate and coordinate similar process across sister campuses; engage other stakeholders (industry, govt)
    - plan for ICRA'22 workshop aiming to report progress on these to-do's (or better versions)

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab

Penn Symposium on Social Implications of Autonomous Military Systems, May 24, 2021

Toward Roboticsts' Ethics for Robots' Design and Use

D. E. Koditschek
ESE Department, University of Pennsylvania
Philadelphia, PA 19104

Penn Engineering | GRASP Laboratory
General Robotics, Automation, Sensing & Perception Lab