Artificial Intelligence – Some Ethical Questions

Professor Bradley J. Strawser
Defense Analysis Department, Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

BACKGROUND:

- *Mala in se vs. Mala in prohibita*
- Technology in cultural context -- De Vries and matchbooks. St. Bernard and Mills.

DISTINCTION: Contingent (technical / empirical) vs. Non-Contingent Problems

MAJOR QUESTIONS:

- **Reasons and Morality**
  Intentions matter: Giving Flowers
  The Racist Soldier; The Sociopath Soldier; … Something else?

- **Opacity / Unintelligibility Problems**
  1. Algorithms can be unfair in ways that cannot be anticipated
  2. Opacity leaves it unclear whether informed consent possible for processes using these algorithms. (*Also connected to command responsibility dilemmas.)
  3. Trusting that which is smarter than us? (Beyond means of verification?)
  4. Many associated challenges. Example: Predictive Policing

  "Citizens targeted by additional police presence (arguably) have a right to know if the targeting mechanism is discriminatory in ways that disfavor them. If the justifiability of the intervention depends on the endorsement of the intervention by its intended beneficiaries, the opacity of the algorithm presents an obstacle to securing endorsement by beneficiaries."

- **Responsibility Problems / So-called Responsibility Dilemma**
  What does it mean to be morally responsible? Agency? Autonomy?
  If agency means X, then AI count as agent. If not, then unclear why important.
  Yet deeper problems concerning category errors of human value/existence

- **Anthropomorphize tools?**
  Are we making a simple category mistake?
  Electric Fences and "Meaningful human control"

- **Game theoretic / competition / adversary problems**
  We may believe restricting AI development, or control, is good or even morally required.
  Adversaries may disagree. That's the "easy" problem.
  The hard problem is that there may be no going back.

- **Moral *obligation* to use Lethal AI?**
  Some threshold of human error in contrast to non-error that makes using human controlled instruments of war irresponsible and impermissible

- **Lowering the Threshold Problem**
  A deep worry; runs far beyond AI
  But perhaps at its greatest zenith; and maybe even different in kind in AI case