Artificial Intelligence: Implications for Autonomous Weapons

Stuart Russell

University of California, Berkeley
Outline

- Remit [etc]
- AI in the context of autonomous weapons
- State of the Art
- Likely future developments
- Conclusions
What is AI?

- The intelligent connection of perception to action
- An AI system is intelligent to the extent that it does the right thing given the available information
- The right thing = the action expected to achieve the goal (or maximize expected utility)
A chess program perceives the current board state and chooses a move

Method 1 (fully autonomous):
- Program considers many future move sequences
- Picks move that wins (or improves material and position structure)

Method 2 (not very autonomous):
- Human provides a table giving the right move for every possible board state
Locus of decision making

- For method 1, human just sets the goal
  - Usually “win the game”
  - Could add “avoid unnecessary captures”
  - Doesn’t say which pieces to capture when
- Machine errors occur due to
  - Limited lookahead
  - Limited ability to take situation-specific features into account when evaluation position
- For method 2, human makes all decisions
- Method 2 is impossible: $10^{55}$ states!
  - Instead, write rules using limited vocabulary of predicates
Partial observability ("fog of war")
Continuous space and time
Uncertainty in action outcomes
Open-universe uncertainty ("unknown unknowns"™):
  objects, behaviors, circumstances unknown at design time
AI: State of the art

- Very rapid progress in recent years
- Dramatically increased industry research
- Much faster computers, much more data
- Sound theoretical framework supports cumulative results and tech transfer
Deep convolutional neural networks achieve human-level object recognition on 1000 categories
  - Note: “Combatant” is not a visual category!
- Face and gait recognition exceed human performance
- Robust, real-time object tracking
- Asynchronous video cameras ~1000fps, low power
Navigation and mapping

- Navigation easy in known environments (e.g., cruise missiles)
- Robot SLAM allows real-time exploration and 3D mapping of city/building/cave
- Autonomous vehicle technology (e.g., MobilEye chip) supports real-time navigation in urban environments with fixed and moving obstacles
Flight control

- Superhuman precision and dynamic control of aerobatic maneuvers, even in tight spaces
- Stabilized platforms for automated snipers
- Insect-sized flapping-wing vehicles
Tactical/strategic decisions

- Entirely self-taught superhuman performance on wide range of “tactical” video games
- Coordinated attack planning in strategy games, RoboCup soccer
- Long-range hierarchical planning in strategy games, logistics
- Combined task and motion planning in robotic applications
- Partial programs provide a continuum from fully preprogrammed to autonomous
Future developments: Timeline

- **Received wisdom: 20-30 years away**
  - “20 or 30 years away from being even possible”
    (techthefuture.com)
  - “could be developed within 20 to 30 years”
    (20 Nobel peace prize laureates)
  - “could become a reality within 20 or 30 years”
    (HRW)
  - “deployment of such systems is, indeed, far off”
    (Anderson and Waxman)

- **On the other hand:**
  - “may come to fruition sooner than we realize”
    (Horowitz and Scharre, 2015)
  - “probably feasible now”
    (UK MoD)
Capabilities

- With work on engineering, systems integration, testing, refinement:
  - Distributed situation awareness
  - Integrated strategic, tactical planning and execution for extended tasks
    - E.g., clear an underground complex
    - E.g., interdict air or ground infiltration over large area
  - Sense/act decision cycles in milliseconds
Physical limits

- The ability of autonomous systems to achieve military goals is likely to be limited more by physics and countermeasures than by AI.
- Limits on range, speed, acceleration, payload, stability, etc., relative to properties of targets.
Physical limits contd.

- Range of form factors from bombers to insect-size
- Low-mass platforms
  - very small caliber projectiles
    - shoot moving human through eyeballs
  - insect-size robots using ~1g shaped charge on contact
- Terminator robots are embarrassingly slow, inaccurate, ineffective
A false debate

- Could an AI drone pilot do a better job of respecting IHL than a human drone pilot, all other things being equal?
- Wrong question! Instead….
- What is the likely end point of an arms race in autonomous weapons, and is that desirable for the human race?
Conclusions

- Sensory, computational, and power limitations are rapidly disappearing
- Core AI capabilities (perception, navigation, mapping, tactical and strategic planning) are, or soon will be in place
- Autonomous systems will be limited by physics more than by AI
- Humans will be largely defenseless against such systems
- Military power will depend much less on manpower, more on money and willingness to deploy
Evolving Position of AI

- Traditional AI position:
  - “Cool toys, lots of funding!”
  - “We take no position on moral or political issues.”

- Current AI position (unofficial):
  - Taints reputation of AI as a field
  - Many oppose such uses of their work

- Debate held at AAAI 15 conference

- Next steps:
  - Written arguments for AI magazine
  - Membership vote on official AAAI policy