Artificial Intelligence: Implications for Autonomous Weapons

Stuart Russell
University of California, Berkeley
Outline

- AI and autonomy
- 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 robot: No autonomy

Take queen!

OK boss
A chess robot: Some autonomy

- take queen
- take rook
- move king
- take rook
- take queen
- move king

really?
A chess robot: Full autonomy

Win!
Locus of decision making

- Even for chess, the human cannot provide, in advance, an individual response for each of $10^{55}$ scenarios.
- Delegating to a machine results in different decisions from those the human would make.
- This is true regardless of whether it’s done by human-defined rules or purely autonomous deliberation.
From chess to warfare
From chess to warfare

- Many more agents and objects
- Partial observability ("fog of war")
- Continuous space and time
- Uncertainty in action outcomes
- Unknown unknowns™

- A theoretical framework is in place to support the development of precise definitions when necessary
AI: Rapidly increasing...

- Computer power
- Data resources
- Industry research
- Capabilities...
“Deep learning” achieves human-level object recognition on 1000 categories
  - Note: “Combatant” is not a visual category!

- Face recognition exceeds human performance
- Vision and 3D sensors track moving objects in real time
- Asynchronous video cameras ~1000fps support high-speed decision cycle
Navigation and mapping

- Navigation is easy in known environments (e.g., cruise missiles)
- Autonomous driving in complex urban environments is harder
- Real-time robot exploration and 3D mapping of city/building/cave
Motor control

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

- **Tactical pursuit/targeting/attack**: entirely self-taught superhuman performance on wide range of video games
- **Coordinated attack planning** in strategy games, RoboCup soccer
- **Long-range hierarchical planning** in strategy games, logistics, construction
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)

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

- Distributed situation awareness
- Integrated strategic, tactical planning and execution for extended tasks, e.g.,
  - Clear an underground complex
  - Prevent ground infiltration over large area
- Sense/act decision cycles in milliseconds
Physical limits

- Systems will be constrained by physics (range, speed, acceleration, payload, stability, etc.), not by AI capabilities
- E.g., lethality of very low-mass platforms is limited by physical robustness of humans:
  - Could use very small caliber weapon
    - Human eyeballs may be the easiest target
  - Could use ~1g shaped charge on direct contact

1g HMTD, 9mm mild steel plate

- [Image of bullet hole]
The IHL matchup: A false debate versus
What is the likely end point of an arms race, and is that desirable for the human race?

Given long-term concerns about controllability of human-level AI systems, should we arm them and turn over our defense to them?

What does a “Flash Crash” look like in the defense arena?
Conclusions

- Sensory, computational, and power limitations are rapidly disappearing.
- Core AI capabilities (perception, navigation, mapping, tactics, strategic planning) are, or soon will be in place.
- LAWS will be limited by physics more than by AI.
- Humans will be largely defenseless.
- Non-state actors could use LAWS to significant effect.
- Controllability will be an increasingly difficult issue.
Evolving Position of AI

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

- Current AI position:
  - “Cool toys, lots of funding!”
  - Autonomous weapons may damage the reputation of AI
  - Professional associations moving towards formal votes on official policy