How To (Not) Trust an Autonomous Weapons System:

The Limits of Trust in Autonomous Weapons Systems

Heather M. Roff
ASU GSI,
Oxford, & New America

David Danks
Philosophy & Psychology
Carnegie Mellon

Autonomous weapons systems

- Focus on machines that can:
  - Identify, select, and engage targets without intervention from a human operator
  - Recognize & respond to shifts in context
  - Construct plans that extend beyond "local" goal
- Planning-autonomy ≠ Learning-autonomy

Probably near-future, not present-day

- Q: When are soldiers able to ***trust*** these AWSs (in the ways required on the battlefield)?

How *not* to think about trust

- From USAF Office of Chief Scientist 2015 report:

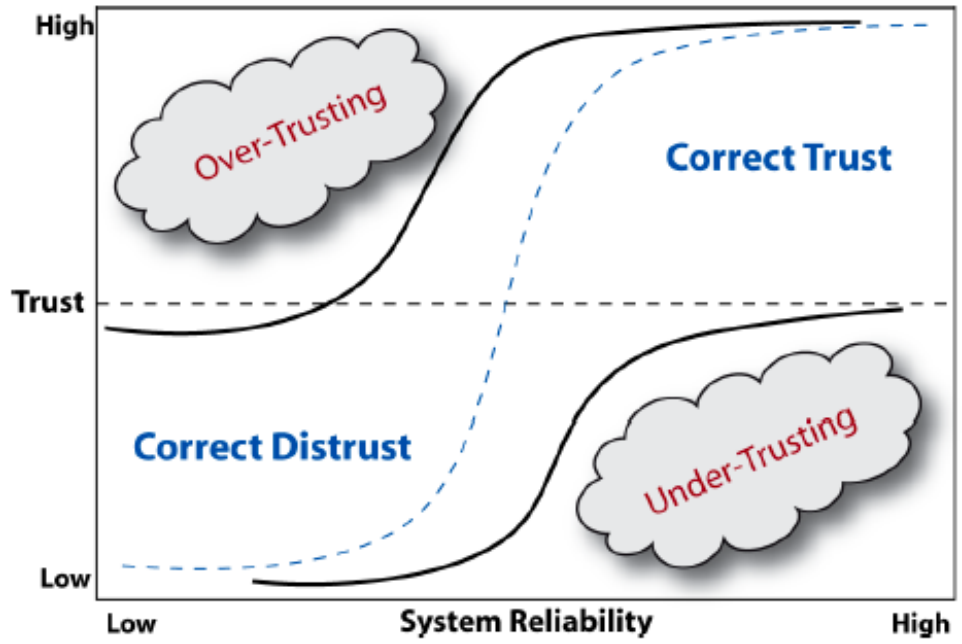

Figure 3. Appropriate calibration of trust in autonomy is critical

Distribution A. Approved for public release; distribution is unlimited. Public Release Case No 2015-0267

Two types of trust

- Predictable & reliable
- Known contexts only
- "Behavioral"
- Know *what* the trustee will do

- Interpretable & sensible
- Novel contexts
- "Cognitive"
- Know *why* the trustee will do

Trust & soldiers

- Soldiers only need to "thin-trust" equipment
  - Equipment need only be predictable & reliable
  - "Why" for weapons, vehicles, etc. is purely mechanistic

- Soldiers must "thick-trust" one another
  - Battlefield involves rapidly changing contexts, and training cannot include them all
  - People really do have 'why's for action

Trust & soldiers: Trust an AWS?

- AWS (of our type) do not cleanly "fit" as either equipment or other soldiers
  - Can exhibit novel behavior in known contexts, and adapt to changing or novel contexts
  - Do not have (human) beliefs, desires, reasons, etc.

- Not mere tools, but not moral agents

Trust & soldiers: Trust an AWS?

- ⇒ Significant barriers to their use (given current training, doctrine, & practice)
  - If AWS is autonomous, then "what" is hard to achieve
  - Soldiers are unlikely to have knowledge of "why"s
    - Our conceptual scheme of reasons, beliefs, desires, etc. need not apply to AWS, particularly learning ones
- Paradoxically, more autonomy ⇒ less trust
  - At least, for near-future AWSs

Possible routes to trust

- Specialized soldier “handler”
  - Analogous with non-human animal “soldiers”
- Integrated training with full unit
  - Boot camp or low-stakes missions
- New development & acquisition process
  - So everyone understands the AWSs & their uses

- Not mutually exclusive! Perhaps need all three?

Thanks!