Policy and Ethics of Intelligent Autonomous Systems

for
Technical Exchange Meeting

Mar 2-4, 2021

Jason Stack

Technical Director
Office of Naval Research
jason.stack@navy.mil

Approved for Public Release

IAS Policy and Ethics

Scope of Effort

Scope: Policy & ethics surrounding development and use of IAS

1. Includes policy & ethics but not law
   - Current international and US law is necessary and sufficient
2. Includes all applications of IAS—i.e., includes but not limited to AWS

Overarching Goals

- Fully inform the DoD's ethical and policy positions and way forward concerning IAS
- Equip everyone engaged with the US National Security enterprise with the perspective and information to engage appropriately in the conversation on ethics and IAS

Desired outcomes

- The appropriate clarifications, elaborations, or modifications to relevant policies and/or ethical principles—as they apply to IAS
- A clear, articulated set of foundational principles, perspectives, and guideposts by which participants can effectively engage in the IAS policy and ethics discussions (both internally & externally)
- Guidance for the technology development community
  (e.g., what development is needed to address policy & ethical considerations)

IAS Policy and Ethics

Definitions & Categorization

Definitions

- IAS = Intelligent Autonomous Systems
  - The general confluence of Autonomy, AI, and Unmanned Systems
  - Includes ***all*** applications (i.e., weaponized and non-weaponized), the human-machine team, "inhabited" platforms (e.g., "diver-less" casualty evacuation), etc.
  - Generally *does not* include "AI at rest" (e.g., ATR algorithms in isolation) nor non-embodied systems (e.g., cyber attack / defense algorithms in isolation)

- AWS = Autonomous Weapon Systems
  - A weapon system that, once activated, can select and engage targets without further intervention by a human operator
  - Includes autonomous, semi-autonomous, human-supervised, etc.

Subsets of IAS

1. AWS
   - Law of War, LOAC, DoD 3000.09 all applicable

2. Non-AWS Higher Risk Applications
   - No belligerent intent, yet physical harm / loss of life is a nontrivial concern
   - e.g., autonomous: cars, rescue, critical drug delivery, casualty evacuation, etc.
   - LoW / LOAC type policies & laws ***not*** applicable

3. General IAS Applications
   - Physical harm / loss of life is not a primary concern
   - e.g., autonomous: search, entertainment, logistics, etc.
   - What about other harms (e.g., JAIC's harms taxonomy including infringement on human rights, denial of opportunity or service, or mission failure)

IAS Policy and Ethics

Strategic Communication Problems

Communication Problems to Solve

- Multiple misconceptions across the DoD significantly impeding progress
  - Myth: "DoDD 3000.09 prohibits autonomous weapons"
  - IAS developers are unclear on what they can / can't do

- Opportunity to better align with our allies and international partners
  - Myth: "IHL / responsible policy requires 'human in/on loop' or 'control' "
  - US position = *Appropriate Levels of Human Judgement* and not *Meaningful Human Control*

- Need to engage in public conversation…major concerns include:
  - Myth: IAS will intentionally / unintentionally do bad things
  - Myth: IAS/AWS will lead to accountability gaps
  - Myth / Hollywood's script formula: AWS will initiate hostilities
  - Myth / implicit assumption: Ethics is an inhibitor and not an enabler




These myths / misunderstandings jeopardize the advancement of US views and the responsible use of IAS in warfare

IAS Policy and Ethics

Making the Conversation Concrete

Key Insight

- An abstract conversation on AWS and ethics is fraught with danger
  - What constitutes an AWS is inconsistently understood / envisioned
  - Hollywood's formula is universally known and emotionally stirring

- Grounding the conversation with an actual—but exemplar—system:
  - Removes most of the confusion, misperception, and miscommunication
  - Elucidates issues not previously identified

Notional System and Situation

- Weaponized quadcopter(s) attacking adversary T-72 tank(s) over the horizon
- Titrate the autonomy and consider policy and ethical implications of:
  - Target selection (i.e., ***that*** T-72, one of ***these*** T-72s, ***any*** target with these specific T-72 attributes)
  - Collateral damage / fratricide potential
  - Human's awareness and ability to terminate engagements
  - Machine's ability to learn
  - Etc.

IAS Policy and Ethics

Problem & Tasking

Problem Statement

1. How do we develop, procure, field, integrate, and employ IAS that:
   - Preserve and maximize warfighting effectiveness
   - Remain consistent with the Law of War & DoD Policy
   - Conform to our broader ethical principles

2. What are these “broader ethical principles” and how are they operationalized?

Tasking

1. Develop recommendations for DoD policy clarifications, additions, or elaborations
2. Apply applicable existing ethical principles to IAS and identify any gaps
3. Articulate the needed assumptions, underlying principles, perspectives, and guideposts that will facilitate a productive conversation surrounding policy and ethics for IAS

Open Questions

Question Sets to be Addressed

Groups of Questions to be Answered

- Autonomous Decision Making
  - What are the nature and bounds of decision-making authority we should give our IAS / AWS?
- System classification
  - What (legal / policy) system classification should we give our IAS?
- Ethical principles
  - Which ethical principles apply to IAS?
  - How do they apply?
- Ethics as an enabler
  - How do we ensure our legal and ethical principles enhance our competitive advantage?

QUESTION SET DETAIL

Proposed Question Set #1

Autonomous Decision Making

What is the nature and bounds of decision-making authority we should give our IAS?

- For affecting itself, it is considerable and understood
  - e.g., sensor settings, propulsion optimization, etc.
- For affecting maneuver, it is measured but understood
  - e.g., path planning, trajectory optimization, etc.
- For affecting the world--in particular strike--it is an open question
  - e.g., can it strike *that* target, *any one of those* particular targets, *any* target with particular attributes, etc.

(Note this question is intentionally about “decision-making authority”—a question has been raised if we should also consider responsibility where responsibility is an obligation to perform a duty)

Action

- Sharpen the question

Proposed Question Set #1

Autonomous Decision Making

What is the nature and bounds of decision-making authority we should give our Autonomous Weapon Systems (AWS)?

- Answer = "It depends." OK, depends on what:
  a) What information the AWS has and its accuracy / staleness (e.g., spatial / temporal bounds, etc.)
  b) Performance (e.g., P_{dc}, P_{fa}, predictable, reliable, cyber safe, etc.)
     – Includes assurance e.g., V&V, T&E, learning in situ, etc.
  c) Human situational awareness and ability to terminate engagements
  d) Collateral damage / fratricide potential
  e) Nature of target (i.e., human vs material) & offensive vs. defensive
  f) Nature and quantity of AWS's available actions (e.g., lethal vs non-lethal)
  g) Ability to ensure appropriate level of reliance (e.g., trust, cognitive overload, etc.)
  h) Ability to avoid unintended escalation
  i) Applicable legal and policy frameworks (e.g. LoW and ROE)
  j) Type of conflict and mission (e.g. ISR, targeting)

Action

- Answer Question 1 wrt dimensions a – j

Need feedback loop to continuously ensure this complex decision boundary is sufficiently clear and in the correct place

Proposed Question Set #1

Autonomous Decision Making

Corollary #1—Using Machines Ethically vs Ethically-Compliant Machines

- Are there a minimum set of "safety features" that prevent unethical employment of IAS or AWS?
  - e.g., ABS for automobiles or GCAS for airplanes
  - Should machines be capable of recognizing unethical actions / orders? What sort of subsequent actions would be appropriate?

Corollary #2—What Changes for (non-AWS) Higher Risk Apps

- Are there new / different policy or legal considerations for (non-weapon) life saving or life preserving IAS applications?
  - e.g., CASEVAC, rescue, etc.
- Is this equivalent to seeking assurance of Asimov's 1st Law (i.e., a robot may not injure or allow harm to a human)

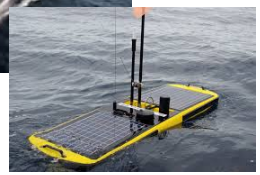
- DANGER: Beware of devolving this question into attempts to solve the Trolley Problem (i.e., a moral dilemma with no solution).

Proposed Question Set #2

IAS "System Classification"

What (legal) system classification should we give our IAS?

- Considerations:
  - Largest to smallest
  - Intent (e.g., intentionally belligerent, env sensing, non-lethal defensive, etc.)



- System Classification examples:
  - Warships, other combatants, military device, etc.
  - US Flagged or "un-flagged"
  - Sovereign immunity or not, etc.
  - Air vs Surface communities
  - Do we need new categories?

- How do we account for:
  - Accountability, responsibility, liability

- Advantages of "higher" classification
  - Rights of state property and combatants (e.g., nav rights, etc.)

- Advantages of "lower" classification
  - Can take more operational risk with IAS, attritable systems, etc.

Action

- Answer wrt all considerations vs. overall warfighting effectiveness

Proposed Question Set #3

IAS Ethical Principles

How do Ethical Principles Apply to IAS?

DoD AI Ethical Principles
(adopted by DoD Feb 21, 2020)

- **Responsible**
- **Equitable**
- **Traceable**
- **Reliable**
- **Governable**

Existing Principles

- DoD AI Ethical Principles
- Asimov / Murphy & Woods Laws of Robotics
- EPSRC Principles of Robotics
- etc.

Open Questions

- How do these principles apply to IAS?
  - Do they apply identically for IAS (including AWS) and AI-enabled systems?
  - If not, what's different?

Approach

- It is generally agreed that principles such as the AI Ethical Principles fully apply and **are necessary** for DoD IAS
- Approach
  - Operationalize the DoD AI Ethical Principles for IAS & investigate sufficiency
  - Examine the field of additional robotics & autonomy principles for applicability and sufficiency to DoD IAS
  - Document identified gaps

Proposed Question Set #3

IAS Ethical Principles

Prior Consideration of AWS

- Prior work has concluded there is no gap in laws or policy re AWS; however, it has suggested there are policy clarifications and elaborations needed
- Additional arguments have been made for the need to consolidate and amplify existing policies ISO debunking myths and misconceptions
- Discussion: Are these "ethical principles", policy positions, guidelines, etc.?

Law of War

DoDD 3000.09

Ethical Principles

- AI Ethical Principles
- et al

Amplifications

- AWS:
  - Are not prohibited
  - Must be used in compliance with LoW
  - Must facilitate "appropriate levels of human judgement"

Clarifications / Derivations

- "Appropriate levels of human judgement" ≠ "human in/on the loop"
- AWS may adjudicate targets within human-selected classes

Elaborations (Obligations on Humans)

- LoW compliance falls on humans (not the AWS)
- Humans must have sufficient info/understanding to determine lawfulness
- Operators must take appropriate action if something goes / is going wrong
- Compliance with law, policy, and ethical principles is the responsibility of humans involved at all stages of development through deployment and implementation

Proposed Question Set #3

IAS Ethical Principles

Ethical principles for IAS

Way Forward

- Pursue previously stated actions (i.e., apply existing principles and analyze for gaps)
- Remain cognizant of the unique strategic communication issues & needs surrounding AWS / lethal autonomous weapons

Corollary #1—Learning in situ

- At what point is in situ learning different (e.g., from “s/w updates”)
- Are there / what are the ethical considerations?
- If the machine is not learning (where it could be), is it actually creating more risk?

Corollary #2—Explainability

- Is *Explainability* critical for DoD IAS or AWS?
  - It is important when machine reasoning is central (e.g., automated loan application approval)—is it central here?

Proposed Question Set #4

Competitive Advantage

How do we ensure our legal and ethical principles enhance our competitive advantage?

- Other actors do not always share our legal and ethical principles
  - These principles prohibit some weapons (e.g., chemical, biological, etc.)
  - These principles govern and constrain most all (e.g., missiles, firearms, etc.)

- Myth? "If our opponents do it and we don't, then we're at a disadvantage"
  - This is sometimes true (e.g., crossbow, gunpowder, aviation, etc.)
  - This is sometimes false (e.g., chemical & biological weapons, etc.)

- Key insights:
  1. Effective countermeasures to a weapon are typically not the weapon itself
  2. Respect for the rule of law is not only "right" but also a competitive advantage—it is an enduring position that humans embrace and often demand
  3. We're not alone: how can we effectively work with allies and partners?

Action

- Understand and articulate how the DoD's legal and ethical principles regarding AWS will impact our warfighting effectiveness vis-à-vis other actors…especially those who do not share our principles

Proposed Question Set #4

Competitive Advantage

Corollary #1—Tactical Effectiveness

- Can IAS enhance operational and/or tactical effectiveness?

Note: Most wartime CIVCAS are LoW compliant and are caused by misidentifications (not collateral damage)

- Argument: Advancements in IAS hold great promise in strengthening compliance with the Law of War, policy, and ethics (e.g., by enhancing the protection of the civilian population against the effects of hostilities)

- Examples:
  - Increasing awareness of civilians and civilian objects on the battlefield
  - Improving assessments of the likely effects of military operations
  - Automating target identification, tracking, selection, and engagement
  - Reducing the need for immediate fires in self-defense
  - Moving from close air support to embedded fires
  - Incorporating “smart” autonomous self-deactivation or self-neutralization mechanisms

Questions

It is not the strongest nor the most intelligent that survives.
It is the one that is the most adaptable to change.

—A core principle from Charles Darwin's
On the Origin of Species

Image from lecture by: Tony Seba, Stanford Univ. Oslo, NOR, Mar 2016

BACKUP SLIDES

IAS Policy and Ethics

Reminder of Select Existing Principles

Law of War (LoW)—Core Principles

- **Military Necessity**: Justifies all lawful measures needed to defeat the enemy (balanced with humanity)
- **Humanity**: Prohibits infliction of suffering, injury or destruction unnecessary to accomplish legitimate military purpose (balanced with necessity)
- **Proportionality**: (jus in bello) prohibition on attacks expected to cause excessive incidental harm. Also underlies requirement to take feasible precautions.
- **Distinction**: Requires distinction between armed forces and the civilian population and between unprotected and protected objects
- **Honor / Chivalry**: Demands certain amount of fairness and mutual respect (e.g. no breach of trust)

DoD LoW Manual (Sec. 6.5.9)

- Sec. 6.5.9: Autonomy in weapon systems not prohibited; LoW obligations apply to persons not machines
- Sec. 6.2.1: All new types of weapons will undergo a legal reveiw

DoDD 3000.09

- *Autonomous weapon systems* require review before formal development and another before fielding
- Goals
  - Minimize risk of failures that could lead to unintended engagements
  - Allow operators to exercise **appropriate levels of human judgement** over the use of force

Proposed Question #N

Why AI Ethical Principles

Why are our ethical principles important?

- Positions the US to ***establish global norms*** for responsible design, development, and use of AI in defense, reflecting our democratic values and strengthening our national interests
- ***Strengthens international partnerships*** by collaborating with allies that share our values and approach to the design, development and use of AI
- ***Earns the trust*** of the American public, industry, and the broader AI community in our focus and commitment to advancing responsible AI
- ***Attracts and retains a competitive digital workforce*** who are critical to preserving and expanding our competitive advantages in AI

Proposed Question #N

Why AI Ethical Principles

DoD AI Ethical Principles

1. **RESPONSIBLE.** DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

2. **EQUITABLE.** The Department will take deliberate steps to minimize unintended bias in AI capabilities.

3. **TRACEABLE.** The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

4. **RELIABLE.** The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

5. **GOVERNABLE.** The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

Backup Material

Definitions

- Accountability
  - answerable; able to explain and justify
  - cannot be shared
- Responsibility
  - executes; a duty to respond and complete
  - can be shared
- Liability
  - legal accountability plus possibility of sanction

- Morals
  - What individuals believe concerning right and wrong
- Ethics
  - Rules for behavior that societies, groups, professions, etc. decide will be followed