Ethical Control of Autonomy

Maritime Risk Symposium 2023, Panel 2

Maritime College, State University of New York

Don Brutzman

Naval Postgraduate School (NPS)

Monterey California

Everything new still must coexist with people…

AI is not magic, and must be tested to be trusted.

Humans remain responsible and in charge.

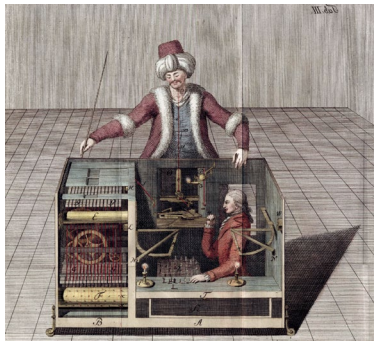
- Otherwise no one is responsible, in charge, liable, safe, etc.
- Robots running amok? Good luck out there!

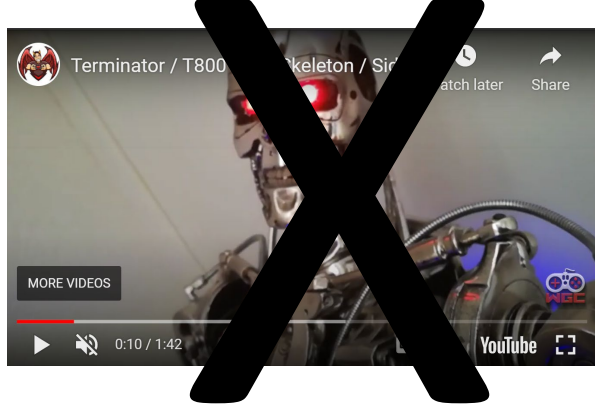
Ethical Control of Autonomous Systems

- Ethical control of autonomous systems can be accomplished through structured mission definitions that are consistently readable, validatable and understandable by humans and robots. Responsible humans must remain in charge of lethal/lifesaving force, and then human-robot teams become more effective.
- https://savage.nps.edu/EthicalControl

ICE BREAKER !!

Awful AI claims are common…





Many people (including big-name AI luminaries) seem to think that some kinda

AI Ethical Agent (perhaps a modern-day Homunculus)

Can sit as a monitoring process on top of any kind of robot software, somehow ensuring that someone else's robot operates morally legally and ethically.

Such misconceptions have inhibited meaningful progress.

(Example: obligatory Terminator image with glowing red eyes.)

Some good work is gong in necessary directions.

Who's in charge? People, software, nothing?

Humans own ethical responsibility and authority.

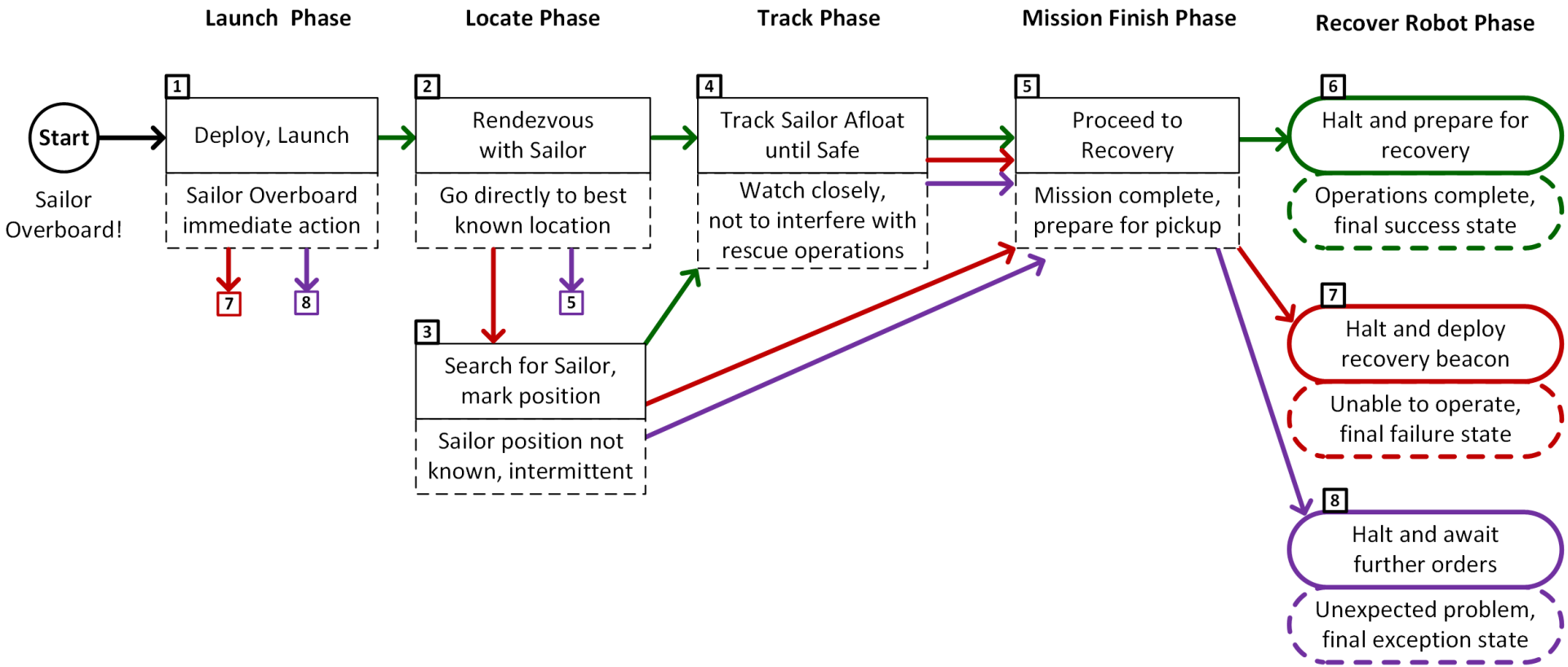
- Legitimacy cannot be fully delegated as ill-defined "autonomy" in AI systems.
- There is no omniscient Delphic-oracle homunculus agent.
- Human-machine combinations can (and perhaps must) be effective

Are we interested in deciding (making moral judgements) whether specific activities are ethical or not?

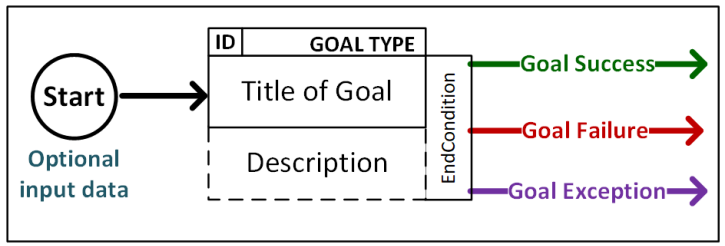
- Responsibility of human actors carrying out shared laws, governance.
- Machines not culpable, legally/ethically/morally, but still can be dangerous.
- In-between case: some human-led organizations avoid responsibility by deferring agency to "autonomous" actors, unsafe/unethical deployment

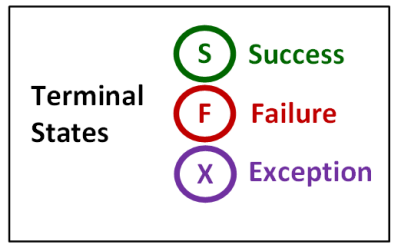
Sailor Overboard, 8 Phases – Mission Execution Automaton (MEA)

Single unmanned air/surface vehicle actions to complement human response when performing "**SAILOR OVERBOARD**" operations, carried out in concert with **shipboard emergency procedures**. Multiple UAVs/USVs can be employed in parallel with ships/aircraft, each following mission orders.



Legend





Don Brutzman and Bob McGhee
Mission upgrade 19 NOV 2019

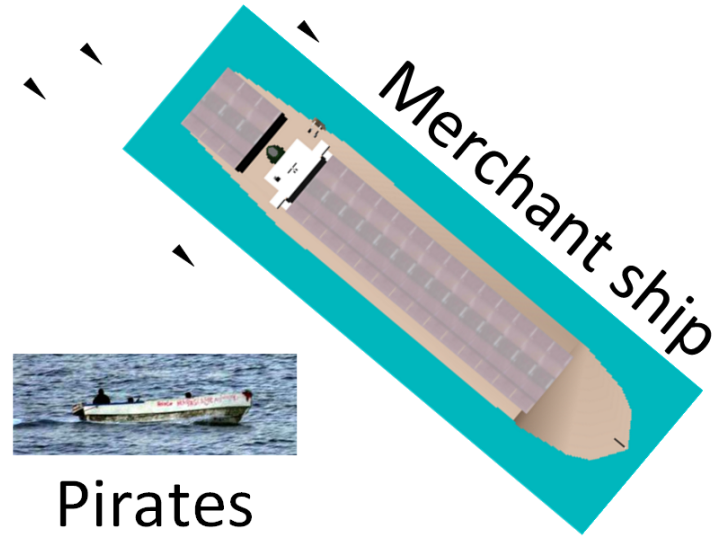
Life boat

Life-saving force: locate, track, communicate, beacon

Ethical control of unmanned systems is required for both lethal and lifesaving force if remote robots communicate intermittently, operating across lengthy time and distance.

Response dilemma for U.S. Navy ship

Lethal force: locate, warn, defend, threaten, attack



Pirates

Respond to one or both scenarios with USV/UAV assets to establish on-scene visibility and presence

OODA Loops for Ethical Control Canonical Missions

Ethical Control OODA Loops	Observe	Orient	Decide	Act
Sailor Overboard	Find Sailor	Report status	Avoid interference	Track sailor until rescued or relieved
Lifeboat Rescue	Find Lifeboat	Report status	Two-way communication	Track life raft until relieved
Pirate Seizure of Merchant Ship	Find merchant ship, pirate small boats	Identity Friend Foe Neutral Unknown (IFFNU) Issue warnings	Human commander authorization to use lethal force	Attack to defend ship if provoked, stay with merchant
Hospital Ship Swarm Attack	EM threat signals detected	(no orientation step in Sense Decide Act)	Reflex-response weapons attack	Mistaken attack on friendly = war crime
Hospital Ship Defense detects spoofing anti-pattern	EM threat signals detected	IFFNU including correlation	Human requirement for lethal force unmet, attack avoided	Report threat alert, commence search for hostile actors

DoD Principles for Ethical AI

Press Release 24 FEB 2020

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.

Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

White House Briefing, 30 OCT 2023

New Standards for AI Safety and Security

- Require that developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government.
- Develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy.

New Standards for AI Safety and Security

Advancing American Leadership Abroad

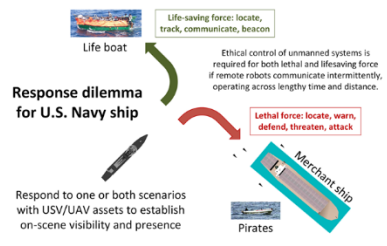
- **Over 2 dozen categories gaining attention…**

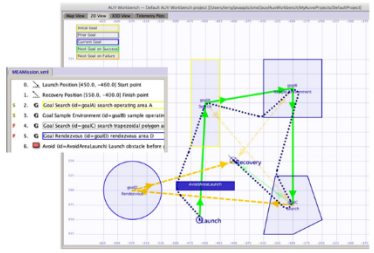
Some takeaways:

- Safety paramount
- Protects economy
- Affects maritime industry at home, influences abroad
- Steps that make sense for human operations are applicable to AI too

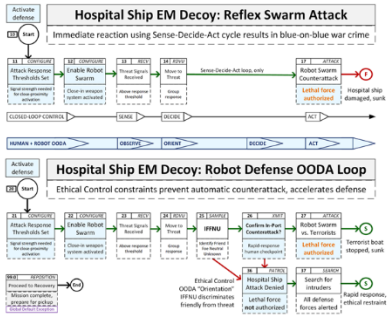
Ethical Control of Unmanned Systems:
Keeping Warfighters in Charge of Autonomy

CRUSER
Consortium for Robotics and Unmanned Systems Education and Research





Mission Execution Ontology (MEO) evolving



Milestones and Transitions

- CRUSER development led to first project selection under CRADA with Raytheon Missile Systems (RMS).
- Successful progress on test missions entering TRL 5 with simulation and Web-sharable 3D visualization.
- Expressing multiple robot mission plans consistently, coherently for diverse UAV, USV, UUV platforms.
- Use Semantic Web Standards to support warfighters.
- Evaluate NAVSEA Unmanned Maritime Autonomy Architecture (UMAA) evolution for robot qualification.

Why / Objectives

- Ethical control of unmanned systems can be accomplished through structured mission definitions that are trusted, consistently readable, validatable, repeatable and understandable by humans and robots.
- Orders must be lawful. Unmanned systems must behave ethically and comprehensibly if they are to support manned military units effectively.
- Well-structured mission orders can be tested and trusted to give human commanders confidence that offboard systems *will do what they are told to do*, and further *will not do what they are forbidden to do*.
- Demonstrate that no technological limitations exist that prevent applying the same kind of ethical constraints on robots and unmanned vehicles that already apply to humans, in lethal and life-saving scenarios.

https://savage.nps.edu/EthicalControl

What / Deliverables

- Update Mission Execution Ontology (MEO) concepts demonstrated in tests and simulation, building to perform field experimentation (FX).
- Supervise thesis work to explore canonical exemplar missions that are expected to utilize unmanned systems, looking across the full range of Naval warfare communities. Example scenarios include UAV for sailor overboard, UAV for refugee/lifeboat escort, and adept scouts. All must observe Law of Armed Conflict (LOAC), Rules of Engagement (ROE), and moral guidance of commanders despite long durations/distances.
- Define, simulate, and test combination of real-world goals and ethical constraints to robot mission tasking across set of canonical scenarios.
- Illustrate how human-robot teams meet moral and legal requirements if deploying unmanned systems with potential for lethal, life-saving force.

NPS PRAESTANTIA PER SCIENTIAM

Building on Concept Demonstrations

Principal Investigator: Don Brutzman brutzman@nps.edu 1.831.656.2149
Co-Investigator: Curtis Blais clblais@nps.edu 1.831.656.3215

Contact

Don Brutzman

brutzman@nps.edu
http://faculty.nps.edu/brutzman

Code USW/Br, Naval Postgraduate School
Monterey California 93943-5000 USA
1.831.656.2149 work
1.831.402.4809 cell